NOWY
SUROWIEC

NOWY SUROWIEC

OTWARTE **ZASOBY DANYCH**
DLA POLSKIEJ GOSPODARKI

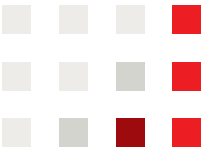
MINISTERSTWO
ROZWOJU

RAPORT

MINISTERSTWO ROZWOJU

NOWY SUROWIEC

OTWARTE **ZASOBY DANYCH** DLA POLSKIEJ GOSPODARKI



autorzy raportu
Krzysztof Węcel
Witold Abramowicz
Piotr Kałużny
Elżbieta Lewańska
Włodzimierz Lewoniewski
Szymon Wieczorek

okładka, projekt, produkcja
Piotr Perzyna: BBW sp. z o.o
NOWE**MEDIA**24.PL

MINISTERSTWO
ROZWOJU

Pl. Trzech Krzyży 3/5
00-507 Warszawa
NIP: 701 079 79 20
REGON: 36 92 67 361
E-mail: kancelaria.mpit@mpit.gov.pl
Tel.: +48 22 262 90 00

Spis teści

1 Wstęp

1.1. Motywacja

Działalność biznesowa coraz mniej opiera się na tradycyjnych czynnikach produkcji, takich jak kapitał czy praca, a w coraz większym stopniu wartość czerpana jest z posiadania i umiejętnego przetwarzania danych. Najwyżej wyceniane firmy na świecie nie posiadają majątku w postaci materialnej: Airbnb jest największym dostawcą miejsc noclegowych, ale nie posiada nieruchomości; Uber jest największą firmą świadczącą usługi przewozowe, ale nie ma żadnego pojazdu. Często sama idea czy obietnica przyszłych zysków może być wyceniana wyżej niż idee dobrze już znane, czego doskonałym przykładem może być Tesla (rys. **1.1** pokazuje porównanie wyceny rynkowej Tesli w porównaniu do GM i Forda).

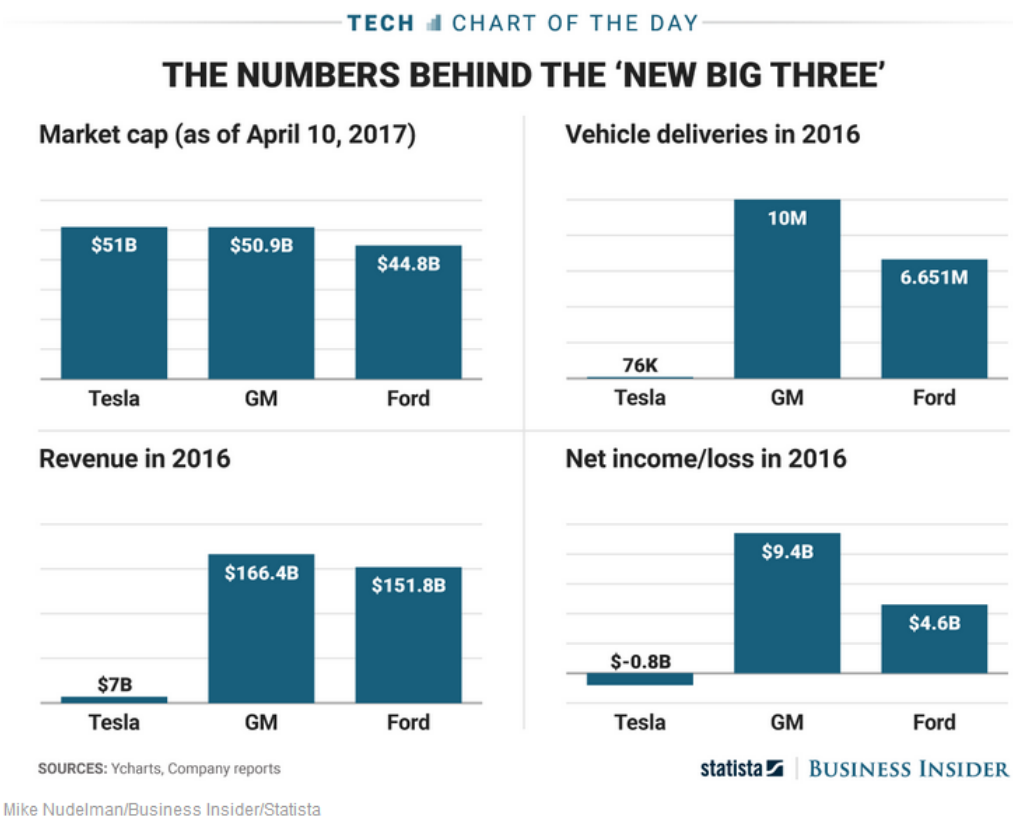
Dostęp do danych jest obecnie jednym z bardziej istotnych czynników wpływających na rozwój przedsiębiorstw. Oprócz tego istotna jest umiejętność ich wykorzystania, a do tego niezbędne jest budowanie kompetencji w zakresie *data science*. Na fali modnego trendu otwierania danych wiele organizacji rządowych, a także firm, udostępnia swoje dane w Internecie. Potencjalni konsumenci danych mają jednak problem ze znajdowaniem właściwych zbiorów, a szczególnie z określaniem ich użyteczności. Wiarygodne i dobrej jakości dane są niezbędne dla budowania modeli mogących wspierać procesy realizowane w przedsiębiorstwach – od produkcji, poprzez marketing do sprzedaży.

Poziom dostępu do danych w Polsce należy określić jako niezadowalający. Można to określić między innymi poprzez badanie dojrzałości cyfrowej przedsiębiorstw. W raporcie o innowacjach opartych na danych [**184**] wykorzystano 29 różnych miar dla określenia poziomu zaawansowania w 28 krajach Unii Europejskiej. Zmienne były zestandaryzowane oraz przeliczone na przedział 0-100 tak, aby zapewnić porównywalność. W rankingu ogólnym Polska zajęła 20. miejsce z liczbą punktów 32,68 (zob. rysunek **1.2**). Najgorszy wynik uzyskał Cypr – 26,88 pkt., natomiast najlepszym krajem okazała się Dania z 71,14 pkt [**184**, str. 12].

Jednym z mierzonych obszarów jest dostępność danych. Aby określić, w jakim stopniu poszczególne państwa udostępniają różne typy danych, określonych zostało 5 wskaźników: gospodarka danymi, otwarte dane, współdzielenie danych, swobodny dostęp do informacji, ochrona wolności słowa. Co ciekawe, zauważono, że nie ma korelacji między poziomem PKB a dostępnością danych w badanych krajach (współczynnik korelacji na poziomie -0.05) [**184**]. Nie ma więc przeszkód, aby kraje uważane za biedniejsze szeroko udostępniały dane. W kategorii „Dane" Polska zajęła 19. miejsce, w kategorii „Gospodarka cyfrowa" – 14. miejsce, „Dostęp do informacji" – 10. miejsce, „Wolność słowa" – 12. miejsce. Dużo słabiej wyglądała sytuacja w kategorii „Danych otwartych" – Polska zajęła 21. miejsce na 22 sklasyfikowane kraje (nie ze wszystkich dane były dostępne), z bardzo niską wartością wskaźnika 12,23 pkt. W ramach open data trochę lepiej wyglądała sytuacja w podkategorii „Oddziaływanie" – 13. miejsce niż „Implementacja" – 22. miejsce[1].

1 Kategoria 'Open Data' została opracowana na podstawie Open Data Barometer, opisywanym w sekcji **3.3.6**

RYSUNEK 1.1. **WYCENA RYNKOWA FIRM SAMOCHODOWYCH: TESLA, GM, FORD**



Źródło: Mike Nudelman/Business Insider/Statista, https://www.businessinsider.de/teslavalue-vs-ford--gm-chart-2017-4

RYSUNEK 1.2. **RANKING OGÓLNY KRAJÓW UNII EUROPEJSKIEJ W ZAKRESIE WYKORZYSTANIA DANYCH DO INNOWACJI – ROK 2017**

RANK	COUNTRY	SCORE	RANK	COUNTRY	SCORE
1	Denmark	71.14	15	Lithuania	43.69
2	Finland	69.36	16	Portugal	39.06
3	Netherlands	65.82	17	Slovenia	37.42
4	Sweden	64.95	18	Latvia	37.22
5	United Kingdom	63.47	19	Slovakia	35.20
6	Estonia	61.11	20	Poland	32.68
7	Austria	53.07	21	Italy	31.29
8	Ireland	49.62	22	Czech Republic	30.80
9	Malta	48.66	23	Romania	30.60
10	Belgium	47.91	24	Greece	28.68
11	France	46.96	25	Croatia	28.10
12	Spain	45.48	26	Hungary	27.46
13	Germany	44.94	27	Bulgaria	26.95
14	Luxembourg	44.47	28	Cyprus	26.88

Źródło: Wallace i Castro [184]

Kolejnym mierzonym obszarem był procent przedsiębiorstw analizujących big data z dowolnego źródła. Jest to oczywiście bardzo ogólna miara, ale pokazuje, dlaczego innowacja oparta na danych ma znaczenie w biznesie. Firmy wykorzystujące rozbudowane modele mogą podejmować lepsze decyzje, a tym samym uzyskiwać przewagę konkurencyjną. W Polsce tylko 5,90% przedsiębiorstw wykorzystuje big data, co daje nam 28. miejsce. Ciekawe jest jednak, że tuż przed Polską uplasowały się Niemcy z wynikiem 5,71%.

Najważniejszą rekomendacją z raportu [**184**] jest konieczność maksymalizowania podaży danych do ponownego wykorzystania. Oznacza to również konieczność zapewnienia swobodnego przepływu danych między przedsiębiorcami. Sam rząd również powinien wykorzystywać dane publiczne do wsparcia własnych procesów.

Polska zajmuje również odległe miejsca według indeksu cyfrowej gospodarki i społeczeństwa DESI (Digital Economy and Society Index)[2]. W dokumencie [**91**] zostało wymienionych 5 wskaźników: Łączność, Kapitał ludzki, Korzystanie z Internetu, Integracja technologii cyfrowej,

2 https://ec.europa.eu/digital-single-market/en/desi

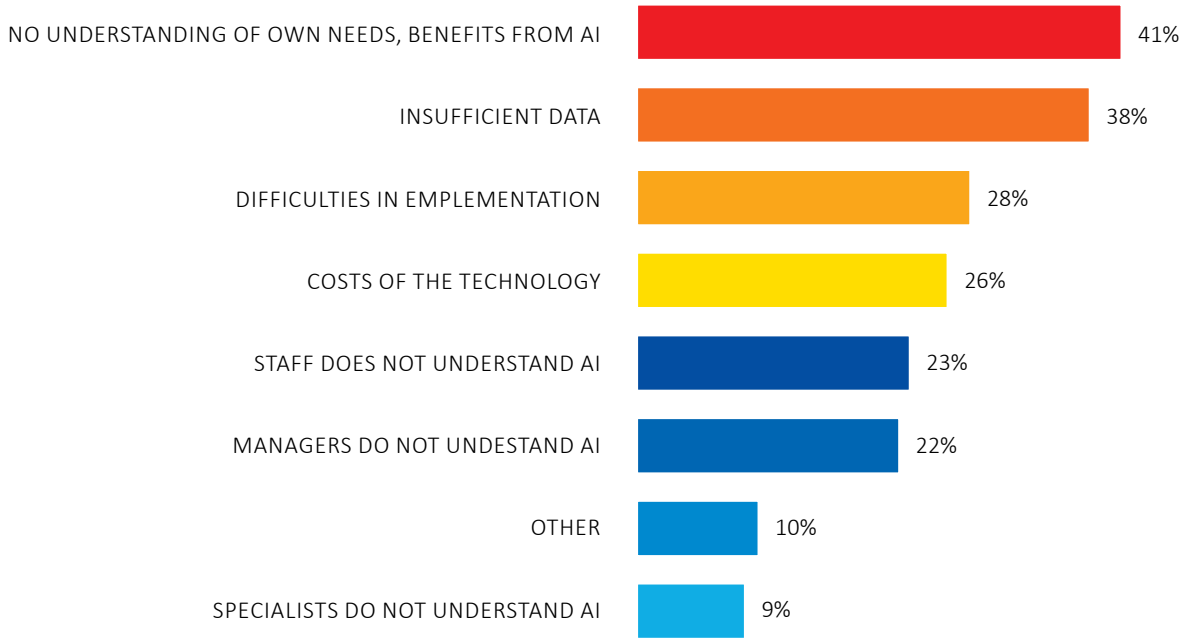
Cyfrowe usługi publiczne, a Polska zajęła w tych kategoriach miejsca odpowiednio: 23., 22., 24., 26., 23. W ogólnym rankingu w badanych latach 2017, 2018 i 2019 utrzymujemy 25. miejsce. Szczególnie słabo wypadają usługi cyfrowe, chociaż z dostępem do internetu jest całkiem nieźle. Rozwój jest widoczny w wielu aspektach, ale czołówka europejska ucieka znacznie szybciej.

Ciekawy wgląd na zapotrzebowanie na dane przedstawia inny raport z mapą rozwoju polskiego AI [**26**], przedstawiony podczas konferencji pt. „Mapa drogowa rozwoju sztucznej inteligencji w Polsce" – MPiT, 22-23.01.2019 r. Należy wziąć pod uwagę, że ankieta, na której opiera się raport, została przeprowadzona wśród firm zajmujących się rozwojem sztucznej inteligencji i nie powinno się uogólniać wyników na wszystkie firmy w Polsce. Niemniej jednak pewne wnioski mogą być interesujące, bo to przede wszystkim firmy z sektora IT są zainteresowane przetwarzaniem dużych ilości danych.

Jeśli chodzi o wykorzystywane dane, to polskie firmy niezbyt często decydują się na ich zakup – robi to zaledwie 31% z nich [**26**]. Z kolei 73% firm przyznało, że wykorzystują dane otwarte (w wolnym dostępie, darmowe). Jako jedną z głównych przeszkód we wdrażaniu AI polskie przedsiębiorstwa high-tech wskazały niedostateczny dostęp do danych – 38% odpowiedzi [**26**]. Była to druga w kolejności najważniejsza bariera rozwoju (por. rys. **1.3**).

Biorąc pod uwagę główne obszary zastosowań sztucznej inteligencji, dostęp do danych jest w istocie krytyczny. Najpopularniejszym zastosowaniem jest przetwarzanie obrazów i rozpoznawanie obiektów – 62% firm wykorzystuje AI w tym obszarze (rys. **1.4**). Inne popularne dziedziny to: eksploracja danych (55%), systemy rekomendujące (52%) oraz przetwarzanie języka naturalnego (43%). To ostatnie jest szczególnie interesujące biorąc pod uwagę zależność od języka – nie można wykorzystać zasobów na przykład dla języka angielskiego, a te są najpopularniejsze. Wszystkie te zastosowania wymagają konkretnego zbioru danych, najczęściej przygotowanego lub wytworzonego przez ludzi.

RYSUNEK 1.3. **PRZESZKODY WE WDRAŻANIU AI W POLSKICH PRZEDSIĘBIORSTWACH**

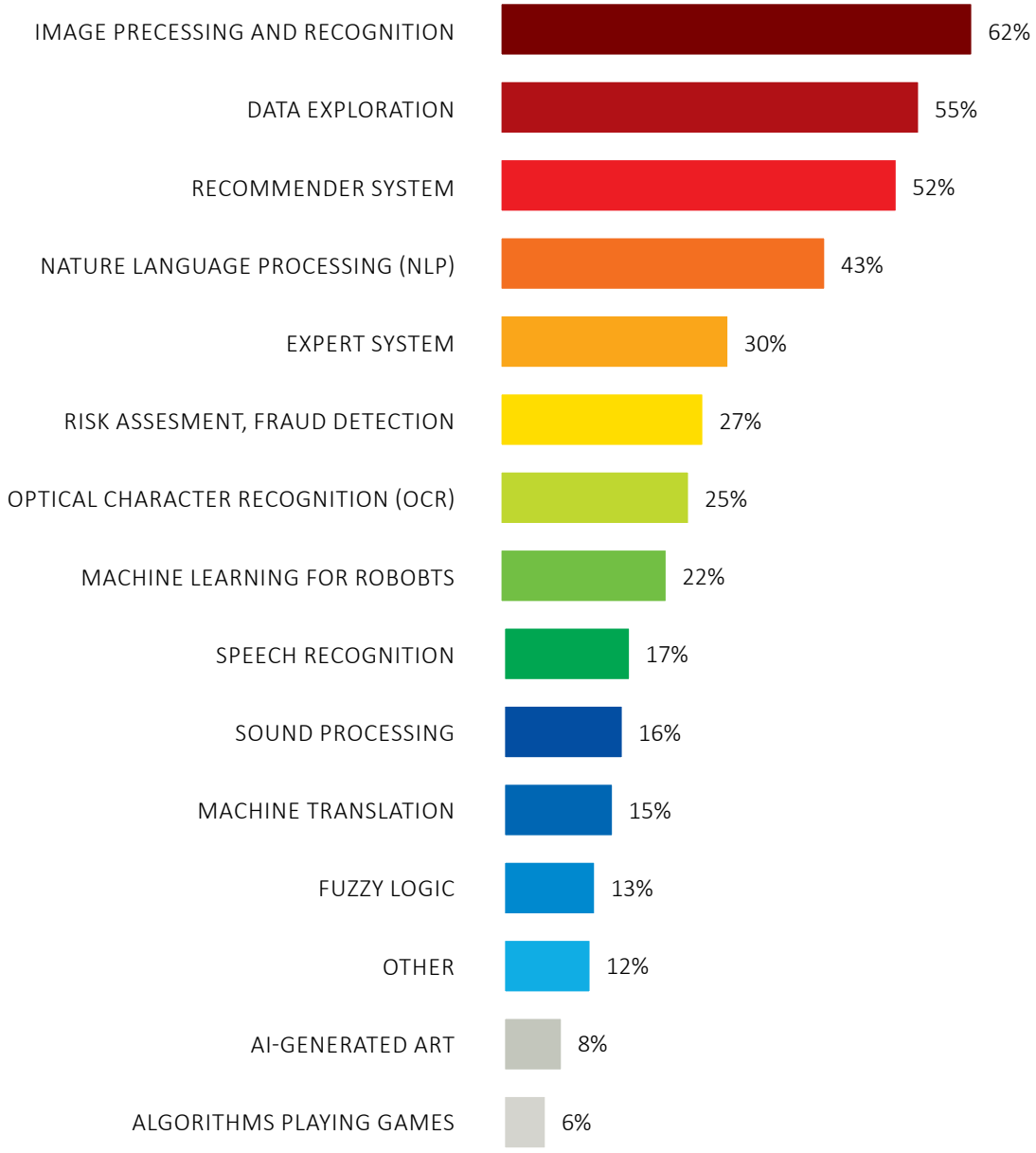


Źródło: [26]

W trakcie wspomnianej konferencji podsumowano również oczekiwania firm wobec państwa [**125**]. Jednym z nich jest otwarcie dostępu do dużych wolumenów danych z administracji publicznej, szczególnie w zakresie rozpoznawania obrazów (RTG/MRI) oraz danych głosowych, w celu rozwoju technologii rozpoznawania mowy dla języka polskiego. W ramach prezentacji MPiT o Narodowej Strategii AI jako jeden z 4 filarów wspomniane zostały dane. W szczególności istotne jest określenie polityki danych, mapowanie zasobów danych oraz infrastruktura do przechowywania i udostępniania danych.

Powyższe raporty wskazują na konieczność podjęcia prac w obszarze zapewnienia lepszego dostępu do danych dla polskich przedsiębiorstw. Przy opracowywaniu polityki pojawia się wiele pytań dotyczących tego zasobu. Celem raportu jest przedstawienie wyników analizy dostępności użytecznych zbiorów danych, które mogą być wykorzystane przez polskie przedsiębiorstwa w celu budowania przewagi konkurencyjnej opartej na danych, w szczególności poprzez rozwój kompetencji w zakresie data science.

RYSUNEK 1.4. **GŁÓWNE ZASTOSOWANIA AI W POLSKICH PRZEDSIĘBIORSTWACH**

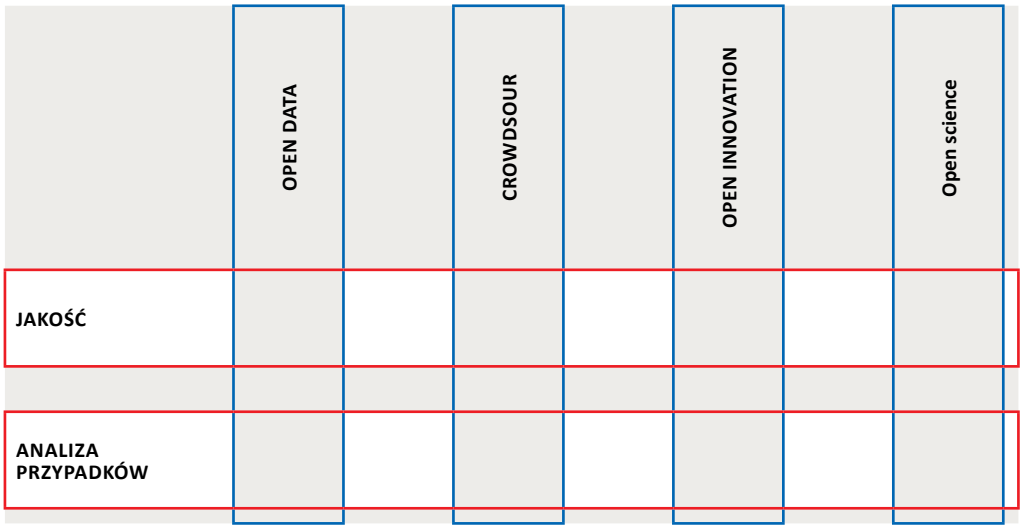


Źródło: [26]

1.2. Metodyka

Podczas prac nad raportem przyjęto następująca metodykę: opisywane zagadnienia zostały podzielone na wertykalne i horyzontalne. Prace były prowadzone równolegle w poszczególnych obszarach. W szczególności w ramach zagadnień wertykalnych wyróżniono następujące kategorie danych: dane publiczne, dane otwarte, dane udostępniane przez firmy oraz dane powstające przy prowadzeniu badań naukowych. Zidentyfikowane zostały również zagadnienia wspólne dla poszczególnych obszarów, które w dalszej części są określane jako horyzontalne. Zagadnienia horyzontalne służą do powiązania powyższych obszarów i obejmują jakość oraz analizy przypadków. Wynikowa struktura prac została przedstawiona na rysunku **1.5**.

RYSUNEK 1.5. **STRUKTURA PRAC NAD RAPORTEM**



Źródło: opracowanie własne.

Jakość danych jest istotna dla każdego zagadnienia horyzontalnego. Odniesienie do jakości pojawia się zarówno w poszczególnych obszarach, jak i jako osobna sekcja w raporcie, gdzie autorzy odnoszą się do powiązań jakości z użytecznością informacji. Druga kwestia to analiza przypadków, które zostały umieszczone w każdym z czterech obszarów.

Przy realizacji ekspertyzy pomocne były bogate materiały źródłowe. Korzystano przede wszystkim z doświadczenia zespołu, z jednej strony w samej dziedzinie danych otwartych (open data) oraz danych powiązanych (linked data), ale z drugiej strony również w zakresie wyszukiwania informacji. Przedmiotem wyszukiwania ze względu na charakter raportu były przede wszystkim zbiory danych. Oprócz wyszukiwania przy pomocy słów kluczowych

zastosowano skuteczną metodę „toczenia kuli śniegowej" (tzw. *snowballing*). Oznacza to, że znalezienie jednego zbioru pozwalało na odnalezienie kolejnych zbiorów. Punktem wyjścia do szczegółowych analiz była pierwotna lista, na której umieszczone zostało ok. 200 linków do zbiorów oraz repozytoriów danych. Poprzez snowballing określono możliwie kompletną kolekcję zbiorów danych, z dodatkowych aspektem weryfikacji przez społeczność – cytuje się dobre źródła, a pomija złe.

1.3. Struktura raportu

Dalsza część raportu zorganizowana została następująco: Rozdział **2** przedstawia definicje danych, pojęć pokrewnych oraz przedstawia rozumienie nasycenia danymi. W rozdziale **3** opisane zostały dane publiczne, czyli w jaki sposób administracja rządowa i samorządowa przyczynia się do zwiększania dostępności danych dla społeczeństwa. Rozdział **4** skupia się na zbiorach tworzonych i dostarczanych przez społeczności; często jako inicjatywy oddolne. O strategiach udostępniania danych przez przedsiębiorstwa można przeczytać w rozdziale **5**. Potencjał uczelni oraz społeczności naukowych w zakresie otwartej nauki został opisany w rozdziale **6**. W każdym z powyższych rozdziałów znalazły się podrozdziały poświęcone analizie przypadków, w których pojawiały się istotne zagadnienia techniczne, np. związane z pozyskiwaniem zbiorów. Rozdział **7** poświęcony jest jakości – zagadnieniu horyzontalnemu, które jest nieodłącznym elementem wszelkich działań związanych z pozyskiwaniem i przetwarzaniem danych. Raport zamyka rozdział **8**, w którym dokonano podsumowania m.in poprzez przedstawienie typologii źródeł oraz rekomendacji.

2 Dane – tematyka raportu

Tematem przewodnim niniejszego raportu są dane, ich źródła i metody przetwarzania. Podstawowym pytaniem w tej sferze pozostaje:

Jakie znaczenie mogą mieć dane i ich wykorzystanie dla administracji oraz przedsiębiorców?

By móc jednak odpowiedzieć na to pytanie, należy najpierw zająć się następującymi zagadnieniami:

1. Opisanie definicji samych danych i ich typologii celem spójnego rozumienia tego pojęcia i wskazania nieklasycznych źródeł danych, które mogą posłużyć do budowania przewagi konkurencyjnej oraz większego poziomu cyfryzacji i informatyzacji działalności prowadzonej przez przedsiębiorców i administrację.
2. Wskazanie praktycznych aspektów definicji i pojęć pomocniczych w stosunku do pierwotnego pojęcia danych, w szczególności uwzględniając objaśnienia dla wielu powszechnie wykorzystywanych pojęć, m.in. big data, open data, nasycenie danymi.
3. Określenie poziomu wykorzystania danych na podstawie przykładów z literatury naukowej, biznesu i raportów administracji – pozwalających określić referencyjny poziom nasycenia danymi dla różnych dziedzin gospodarki.
4. Opracowanie praktycznego przewodnika po źródłach i typach danych, ze wskazaniem celowości niniejszego raportu i jego głównego celu w perspektywie opisu źródeł danych.

2.1. Definicje danych

Dane są pojęciem elementarnym i trudnym w definiowaniu, na co wskazuje brak jednoznacznej definicji. Wielu autorów podejmowało wyzwanie zdefiniowania tego pojęcia, czy to bardziej ogólnego, czy też na potrzeby własnej pracy. Dane mogą reprezentować różnego rodzaju fakty, być odwzorowaniem stanu obiektów (w tym takich jak np. dokumenty) czy też odczytów pochodzących z urządzeń i systemów informatycznych. W perspektywie tego raportu warto skupić się na przedstawieniu tego, jakie źródła danych i w jaki sposób mogą być wykorzystane w celu budowania przewagi konkurencyjnej. Celem tego podrozdziału nie jest zaprezentowanie wyczerpującej klasyfikacji źródeł danych z perspektywy naukowej, a raczej zwrócenie uwagi na szerszy sposób interpretacji pojęcia dane niż zwyczajowo przyjęty. W rozumieniu pojęcia danych w literaturze może pomóc szerokie spektrum definicji przedstawionych w tabeli **2.1**.

TABELA 2.1. **PRZYKŁADY DEFINICJI DANYCH I INFORMACJI**

AUTORZY	DEFINICJE DANYCH	DEFINICJE INFORMACJI
Avison i Fitzgerald (1995) [15]	Dane reprezentują nieustrukturyzowane fakty.	Informacja ma znaczenie... pochodzi z wyselekcjonowania danych, ich podsumowania i prezentacji w taki sposób, by były użyteczne dla odbiorcy.
Clare i Loucopoulos (1987) [31]	Fakty zgromadzone z obserwacji lub zapisów dotyczących zjawisk, obiektów lub ludzi.	Wymagania do podejmowania decyzji. Informacje są produktem istotnego przetwarzania danych.
Galland (1982) [56]	Fakty, koncepcje lub wyniki w postaci, która może być komunikowana i interpretowana.	Informacje to to, co powstaje w wyniku pewnych działań myślowych człowieka (obserwacji, analiz) z sukcesem zastosowanych do danych, by odkryć ich istotę lub znaczenie.
Hicks (1993, 3 rd Ed) [72]	Reprezentacja faktów, koncepcji lub instrukcji w sposób sformalizowany, umożliwiający komunikowanie, interpretację lub przetwarzanie przez ludzi lub urządzenia automatyczne.	Dane przetworzone tak, by miały znaczenie dla decydenta w konkretnej sytuacji decyzyjnej.
Knight and Silk (1990) [89]	Liczby reprezentujące obserwowalne obiekty lub zagadnienia (fakty).	Znaczenie dla człowieka związane z obserwowanymi obiektami i zjawiskami
Laudon and Laudon (1991) [94]	Surowe fakty, które mogą być kształtowane i formowane, by stworzyć informacje	Dane, które zostały ukształtowane lub uformowane przez człowieka w istotną i użyteczną postać.
Maddison (1989) [111]	Podane fakty, z których inni mogą dedukować, wyciągać wnioski. W Informatyce: znaki lub symbole, w szczególności w transmisji w systemach komunikacji i w przetwarzaniu w systemach komputerowych; zwykle choć nie zawsze reprezentujące informacje, ustalone fakty lub wynikającą z nich wiedzę; reprezentowane przez ustalone znaki, kody, zasady konstrukcji i strukturę.	Zrozumiała, użyteczna, adekwatna komunikacja w odpowiednim czasie; jakikolwiek rodzaj wiedzy o rzeczach i koncepcjach w świecie dyskusji, która jest wymieniana pomiędzy użytkownikami; to treść, która ma znaczenie, a nie jej odwzorowanie.
Martin and Powell (1992)	Surowce życia organizacji; składają się z rozłącznych liczb, słów, symboli i sylab odwołujących się do zjawisk i procesów biznesu.	Informacje pochodzą z danych, które zostały przetworzone tak, by stały się użyteczne w podejmowaniu decyzji w zarządzaniu.
Soukhanov (1999) [163]	Dane są przekazywane w formie liczb, tekstów, rysunków, czy dźwięków.	Przez informację rozumiemy kolekcję danych o pewnych obiektach pozwalających na jej komunikowanie, organizowane i prezentowane w usystematyzowany sposób, wyjaśniający ich znaczenie.
Hackathorn (2001) [66]	Dane są wynikiem obserwacji zjawisk, rzeczy i osób.	To dane, które pokazują możliwość podejmowania decyzji lub je powodują. Dane stają się informacjami po zinterpretowaniu ich przez ludzi.

Źródło: [1] oraz tłumaczenie [63] na podstawie [76]

Wspólnym elementem powyższych definicji jest to, że dane opisują fakty, natomiast z przetworzonych danych możemy uzyskać istotne dla odbiorcy informacje. U podstaw każdego procesu budowania wartości z danych jest więc przeobrażenie danych w dobre informacje, które mogą być charakteryzowane przez różnorodne cechy, takie jak:

- **Aktualność** – zdolność do odzwierciedlania rzeczywistości. Informacja jest aktualna jeśli opisuje teraźniejszy stan pewnej rzeczywistości lub ostatni możliwy do rozpoznania (ale wystarczający dla naszych potrzeb). Informacja może być odniesiona do czasu wskazanego w przeszłości, przykładem jest notowanie giełdowe sprzed roku.
- **Kompletność** – uwzględniająca wszystkie istotne cechy, przekazująca jak najprawdziwszy i najpełniejszy obraz opisywanego obiektu bądź zdarzenia. Informacja kompletna to taka, która nie pomija i nie agreguje cech, dając jak najpełniejszy obraz obiektu, który opisuje.
- **Prawdziwość** – informacja jest prawdziwa, jeżeli opisuje w przyjętych jako dopuszczalne granicach błędu stan pewnej rzeczywistości. Może dotyczyć wartości dyskretnych, np. włącznik włączony/wyłączony lub niedyskretnych tj. kolor czy poziom naświetlenia.
- **Wiarygodność** – jest dodatkową miarą prawdziwości i aktualności informacji, określającą, na ile wskazane cechy odzwierciedlają rzeczywisty stan obiektu. Niekiedy nie mogąc określić wiarygodności samej informacji, przyjmuje się, że informacja jest tak wiarygodna jak jej źródło.
- **Użyteczność/relewantność informacji** – ważność, odpowiedniość informacji, która przypisywana jest przez użytkownika. Informacja jest użyteczna, jeżeli odpowiada potrzebom jej odbiorcy. Relewantność jest określana przez odbiorcę, a nie nadawcę informacji. Ta sama informacja może być użyteczna dla jednych, a dla innych może być użyteczna tylko w określonym czasie lub wcale.
- **Przyswajalność informacji** – informacja jest przyswajalna, gdy odbiorca może ją wykorzystać bez konieczności wykonywania dodatkowych operacji przekształcających. By określić tę cechę, można mierzyć np. liczbę kroków, które musimy wykonać, aby uzyskać informację użyteczną.

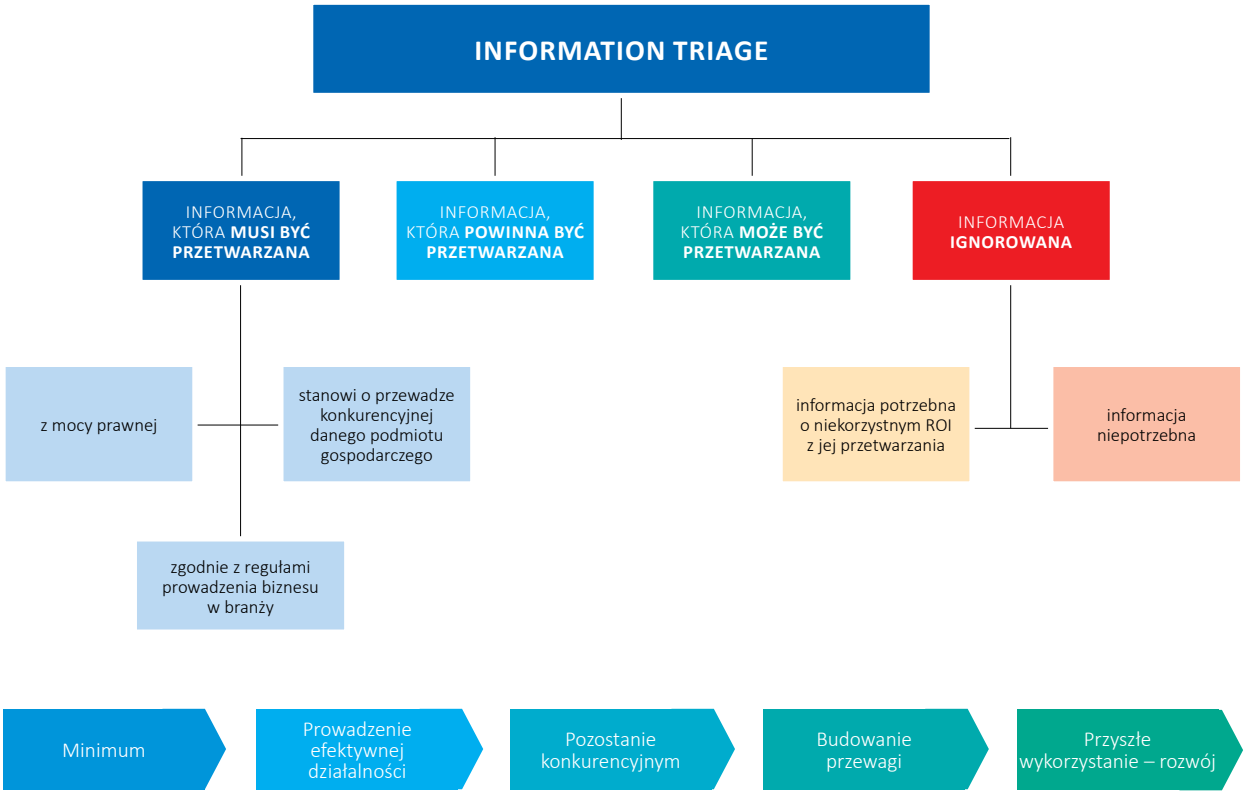
Cechy te mają bardzo praktyczny charakter, ponieważ wskazują na kryteria wartości informacji jako zasobu i możliwości korzystania z niej w realiach organizacji – czy to przedsiębiorstwa, czy jednostki publicznej. Dane jako nośnik odzwierciedlający rzeczywistość posiadają różną wartość dla różnego rodzaju podmiotów wynikającą z różnej użyteczności dla odbiorców. Ich przetworzenie na informacje o pożądanych cechach może służyć do wspomagania lub automatycznego podejmowania decyzji, zgodnie z ideą „information triage", przedstawioną na rysunku **2.1**. Pojęcie to jest kluczowe dla zrozumienia wykorzystania informacji przez podmioty gospodarcze. Zgodnie z tą koncepcją organizacja musi przetwarzać informacje:

- do których przetwarzania jest zobowiązana z mocy prawa, np. dane finansowe, o zatrudnieniu,

- które wynikają z reguł prowadzenia biznesu w danej branży (dane o kontrahentach i klientach, dane sprzedażowe),
- stanowiące dla organizacji unikalny zasób dostępny tylko dla niej, w rozumieniu najbardziej podstawowego źródła przewagi konkurencyjnej.

Te kategorie najbardziej niezbędnych informacji definiują minimalny kanon źródeł danych, które muszą zostać przetworzone przez organizacje do "dobrych" – posiadających wiele wyżej wymienionych cech – informacji. Jednak może to pozwolić co najwyżej na sprawne zarządzaniem przedsiębiorstwem i prowadzenie działalności w "w danym momencie" – czyli kontrolę nad bieżącymi działaniami firmy. Przetwarzanie w sposób efektywny i bezbłędny kategorii informacji, które muszą być przetwarzane, już może dla organizacji być procesem skomplikowanym, ponieważ często wymaga bardzo dobrego zrozumienia działania całej organizacji i wysokiego poziomu cyfryzacji jej zasobów informacyjnych.

RYSUNEK 2.1. **POJĘCIE INFORMATION TRIAGE**



Źródło: Opracowanie na podstawie [1].

Jednak osiągniecie takiego poziomu nie pozwala w tym momencie na konkurowanie z dużymi organizacjami, które korzystają dodatkowo z "nieklasycznych" kategorii danych, do których uzyskały dostęp. Wzbogacanie informacji uzyskanych z podstawowej działalnością organizacji, np. danych sprzedażowych, o dane z samego procesu sprzedaży firmy, tworzy informacje, które powinno się przetwarzać. Pozwalają one ulepszyć podstawową działalność organizacji i pozwolić na jej optymalizację. **Organizacja, która nie ma chęci przetwarzania innych informacji niż niezbędne, może w przyszłości stracić przewagę konkurencyjną, gdy jej konkurencji znajdą źródła pozwalające na wzbogacenie ich pierwotnego strumienia informacji niezbędnych**. W zależności od branży i potrzeb można mówić tutaj o danych o zachowaniu klientów, zapobieganiu nadużyciom, z obserwacji procesów wewnętrznych i badania ich efektywności. By jednak organizacja mogła budować swoją przewagę, musi korzystać także ze źródeł często znajdujących się poza ramami jej bezpośredniego otoczenia, wzbogacając swój asortyment dostępnych informacji o zewnętrzne źródła. Takie informacje, które mogą być przetwarzane, obejmują także informacje możliwe do uzyskania z wykorzystaniem algorytmów AI, nowoczesnych technologii przetwarzania i zewnętrznych źródeł danych otwartych, a także efektywnej wymiany danych z partnerami biznesowymi i organizacjami publicznymi. Końcowym etapem budowania wartości jest także badanie wykorzystania nowych źródeł informacji do budowy usług – nawet takich, które w danym momencie mogą nie przynosić firmie bezpośredniego zysku lub wiązać się z niekorzystnym zwrotem z inwestycji (ROI). To właśnie ciągłe testowanie nowych podejść pozwala organizacji na długofalowy rozwój. Oczywistym wnioskiem z powyższego jest to, że aby mówić w ogóle o budowaniu przewagi konkurencyjnej, najpierw **organizacja musi doskonale poznać dane znajdujące się w jej wnętrzu i te dostępne dla niej w jej najbliższym otoczeniu gospodarczym**.

2.2. Definicje pojęć powiązanych

Obok pojęcia „dane" w powszechnym obiegu występuje wiele pojęć powiązanych. Interpretacja pojęć takich, jak informacja publiczna czy dane (w tym dane otwarte) jest spójna z raportem „Otwieranie danych – podręcznik dobrych praktyk" przygotowanym przez Ministerstwo Cyfryzacji pod koniec listopada 2018 roku [**120**][1]. W celu jednoznacznego rozumienia pojęć wykorzystywanych w raporcie wprowadzamy poniższe definicje:

Big Data [**185**]. Najszerszym do zdefiniowania pojęciem w tej kategorii jest big data (dane wielkoskalowe). Oznacza ono dane nie tylko o dużym wolumenie (miliardy rekordów, a często więcej), ale przede wszystkim różnorodność, zmienność i nietypowy format danych, które wymagają unikalnych kompetencji i narzędzi do ich przetwarzania. Do najważniejszych cech należy zaliczyć (w języku angielskim określa się je literami "V"):

1 Jest to pozycja skierowana przede wszystkim do pracowników administracji rządowej oraz samorządowej – w szczególności odpowiadających za usługi udostępniania danych. Zapoznanie się z nią może być dla tych osób istotne.

- Duża ilość danych (ang. volume) – mówiąca tutaj o przetwarzaniu takiego wolumenu danych, gdzie klasyczne narzędzia analityczne (arkusze, relacyjne bazy danych) wykorzystywane dotychczas w organizacji nie są dostatecznie wydajne lub w ogóle nie działają.
- Duża prędkość przetwarzania danych (ang. velocity) – w rozumieniu prędkości mówi się o dynamice generowania danych, które powstają na bieżąco, w trybie ciągłym, w sposób zautomatyzowany. Wymaga to specyficznego sposobu przetwarzania i analizy, korzystając z narzędzi pozwalających na analizę strumieni danych. Ważna jest tutaj zmiana podejścia, gdzie analiza odbywa się na bieżąco, a nie z danych znajdujących się np. w hurtowni danych. Często oznacza to również dane, które powstają w urządzeniach Internetu Rzeczy. Duża prędkość może oznaczać, że niektóre operacje chcemy wykonać częściej (np. analizę awarii i anomalii), a niektóre mogą być wykonane w późniejszym czasie.
- Duża różnorodność danych (ang. variety) – w przeciwieństwie do znanego relacyjnego modelu, w przypadku big data możemy mieć do czynienia z danymi grafowymi, geograficznymi, tekstem, a także obrazem czy dźwiękiem. Wymaga to skorzystania ze specjalistycznych narzędzi dostosowanych do różnorodnych formatów przetwarzanych danych, a także unikalnych kompetencji ekspertów potrafiących przetwarzać np. dane w postaci obrazu.
- Duża niepewność (ang. veracity) – określana również jako brak precyzji w danych, zwraca uwagę na potrzebę weryfikacji posiadanych danych i określenie ogólnego poziomu ich jakości. Dynamicznie generowane dane o dużym wolumenie mogą zawierać braki bądź błędy. Umiejętność radzenia sobie z taką charakterystyką danych, ich rzetelnego sprawdzenia, zauważenia braków, błędów i ich przyczyn pozostaje jedną z kluczowych kompetencji potrzebnych do przetwarzania takich zbiorów.

Często do przetwarzania zbiorów o takiej charakterystyce stosuje się więc specjalistyczne algorytmy, podejścia, frameworki i narzędzia data science. Do predykcji, klasyfikacji i innego rodzaju wnioskowania na wyżej wymienionych danych stosuje się natomiast często algorytmy związane ze sztuczną inteligencją (AI) – w rozumieniu uczenia maszynowego (ML), a także sieci neuronowych (NN).

Informacja publiczna[2]. Jest to każda informacja o sprawach publicznych. Prawo do informacji publicznej obejmuje uprawnienie do niezwłocznego uzyskania informacji zawierającej aktualną wiedzę o sprawach publicznych. Zobowiązane do udostępniania informacji publicznej są władze publiczne oraz inne podmioty wykonujące zadania publiczne:

a) organy władzy publicznej,
b) organy samorządów gospodarczych i zawodowych,
c) podmioty reprezentujące zgodnie z odrębnymi przepisami Skarb Państwa,

2 Zgodnie z jednolitym tekstem ustawy o dostępie do informacji publicznej [**158**].

d) podmioty reprezentujące państwowe osoby prawne albo osoby prawne samorządu terytorialnego oraz podmioty reprezentujące inne państwowe jednostki organizacyjne albo jednostki organizacyjne samorządu terytorialnego,
e) podmioty reprezentujące inne osoby lub jednostki organizacyjne, które wykonują zadania publiczne lub dysponują majątkiem publicznym, oraz osoby prawne, w których Skarb Państwa, jednostki samorządu terytorialnego lub samorządu gospodarczego albo zawodowego mają pozycję dominującą w rozumieniu przepisów o ochronie konkurencji i konsumentów.

W przypadku informacji publicznej w szczególności, mówić można o:

- polityce wewnętrznej i zagranicznej – w tym projektowaniu ustaw i dokumentach związanych z procesem ich realizacji,
- organizacji, kompetencjach i strukturze podmiotów, a także prowadzonych przez nich działaniach, utrzymywanych rejestrach, stanie rozpatrywanych spraw,
- danych publicznych – treści dokumentów urzędowych, stanowisk i treści publikowanych przez organy władzy publicznej, informacji o stanie państwa i samorządów oraz ich jednostek organizacyjnych,
- majątku publicznym – Skarbu Państwa, państwowych osób prawnych, jednostek samorządowych, długu publicznym, pomocy publicznej.

Dane publiczne[3]**.** Liczby i pojedyncze wydarzenia lub obiekty na możliwie najniższym poziomie agregacji, które nie zostały poddane przez administrację publiczną przetworzeniu do postaci raportów, wykresów itp. oraz nie został im nadany odpowiedni kontekst lub interpretacja; istotne są dwa aspekty: możliwie najniższego poziomu agregacji (np. dla gmin zamiast powiatów) przy zapewnieniu ochrony danych osobowych oraz udostępniania danych, a nie gotowych raportów czy opracowań.

Informacje sektora publicznego (ISP). Każda treść lub jej część, niezależnie od sposobu utrwalenia, szczególnie w postaci papierowej, elektronicznej, dźwiękowej, wizualnej lub audiowizualnej, będąca w posiadaniu podmiotów z sektora publicznego (wymienionych w ustawie o ponownym wykorzystaniu informacji sektora publicznego) [**158**].

Ponowne wykorzystywanie informacji. Wykorzystywanie przez osoby fizyczne, osoby prawne i jednostki organizacyjne nie posiadające osobowości prawnej, informacji sektora publicznego, w celach komercyjnych lub niekomercyjnych innych niż pierwotny cel, dla którego informacja została wytworzona.

3 http://archiwum.mc.gov.pl/projekty/otwartoscdanych/slowniczek-otwartosc-danych

Dane otwarte (open data). O danych w pełni otwartych[4] możemy mówić, gdy zgodnie z podstawową definicją: *są dostępne, a więc każdy może je pobrać i udostępniać, przetwarzać i wykorzystywać, np. dzięki dedykowanym portalom lub usługom, w dowolnych celach, w tym do celów gospodarczych.*

Ważnymi elementami wskazującymi jednak na praktyczny charakter wykorzystania open data w rozumieniu publicznym są kryteria dotyczące:

- **Dostępności / kompletności** – dane mają być dostępne w sposób stały, bez dodatkowych barier i bez wprowadzania uprawnień w celu uprzywilejowania konkretnej grupy osób.
- **Możliwości i efektywności przetwarzania** – dane powinny być udostępnione online, z wykorzystaniem dedykowanych portali i usług. Metody ich publikacji i udostępniania nie powinny uniemożliwiać dostępu bez posiadania specjalistycznego oprogramowania (np. pliki AutoCAD). Dodatkowo dane powinny być udostępnione zgodnie ze standardami i formatami branżowymi oraz technicznymi, co powinno pozwolić na automatyzację dostępu, np. za pomocą API.
- **Agregacji** – dane mają być udostępniane w postaci jak najmniej przetworzonej.
- **Źródła** – dane powinny być zbierane, a źródło powinno być stałe i możliwe do weryfikacji.
- **Aktualności** – dane powinny być udostępniane tak szybko, jak to możliwe (co zakłada w określonych przypadkach istnienie automatycznej usługi je udostępniającej w sposób dynamiczny). Nie powinny występować żadne dodatkowe opóźnienia w tym procesie.
- **Licencji** – powinna być określona licencja, na której dane są udostępnione, zgodna z podstawowymi założeniami open data. Niemożliwe jest zastosowanie żadnych ograniczeń dostępu, patentów, copyrightów, trademarków czy regulacji związanych z tajemnicą gospodarczą w stosunku do udostępnianych danych. Możliwe jest jednak zachowanie staranności w celu zapewnienia odpowiedniej prywatności i bezpieczeństwa zasobu.

Poziomy otwartości Open Data. Aby dokładnie rozpatrywać jakość udostępnionych zbiorów danych, można określić ich zgodność z wyżej wymienionymi cechami. Inny sposób został zaproponowany przez twórcę idei WWW i Linked Data Tima Bernersa-Lee. Określił on 5-poziomowy sposób oceny udostępnianych danych, tzw. poziomy dostępności[5], pokazane na rys. **2.2** oraz w tabeli **2.2**.

4 http://opendefinition.org/, https://blog.okfn.org/2013/10/03/definingopen-data/, https://opengovdata.org/.

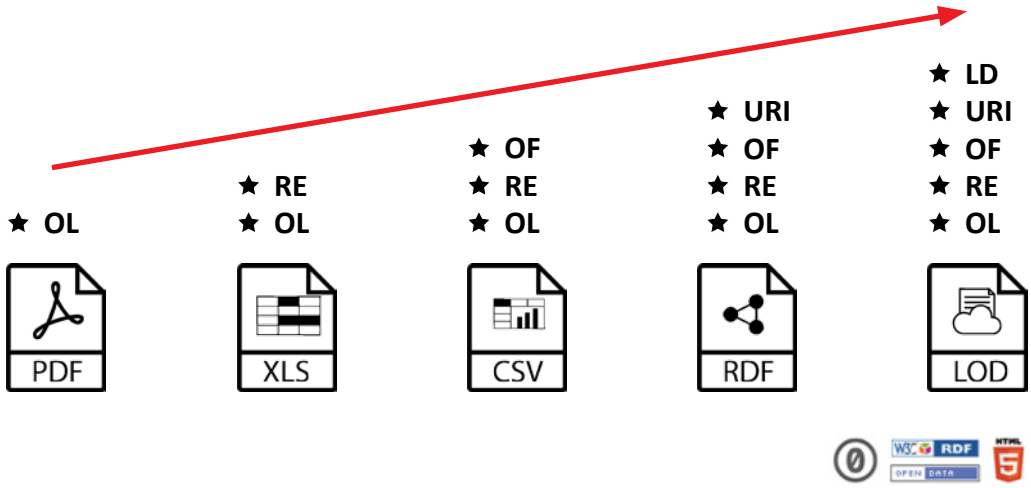
5 https://5stardata.info/en/l

TABELA 2.2. **OCENA POZIOMU OTWARTOŚCI DANYCH NA PODSTAWIE PIĘCIU GWIAZDEK**

POZIOM	OPIS	NAJWIĘKSZA ZALETA	FORMATY
★	Udostępnienie danych w sieci Web (w dowolnym formacie) na warunkach otwartej licencji. Ze względu na późniejsze kryteria są to zwykle pliki w formacie nieprzetwarzanym w sposób automatyczny (zdjęcia lub pliki PDF).	Dostępność danych.	PDF, JPG (skany)
★★	Udostępnienie danych w formie ustrukturyzowanej, co pozwala na ich automatyczne przetwarzanie, mogą wymagać specyficznego licencjonowanego oprogramowania do otwarcia (np. arkusz kalkulacyjny zamiast zeskanowanego obrazu tabeli w przypadku powyżej).	Możliwość automatycznego przetwarzania.	XLS (pliki Excel)
★★★	Używanie formatów otwartych (np. CSV zamiast arkusza kalkulacyjnego).	Niezależność od posiadania licencji programu.	CSV, XML, JSON
★★★★	Używanie URI do oznaczania zasobów, aby możliwe było ich wyszukiwanie i pobieranie. Dane są oznaczone za pomocą identyfikatorów URI/URL – możliwe jest ich wyszukiwanie i automatyczny dostęp, są one zgodne ze standardami W3C.	Możliwość automatyzacji pobierania i wyszukiwania danych – budowanie usług.	RDF i SPARQL
★★★★★	Dane są połączone z innymi zbiorami (Linked Data), co pozwala na uzyskanie kontekstu przetwarzanych danych.	Możliwość łączenia różnych zbiorów.	LOD (zasoby ostępne jako Linked Data)

Źródło: na podstawie https://5stardata.info/en/ oraz [120]

RYSUNEK 2.2. **OCENA OTWARTYCH DANYCH PRZY POMOCY PIĘCIU GWIAZDEK**



Źródło: https://5stardata.info/en/

2.3. Źródła i typy danych

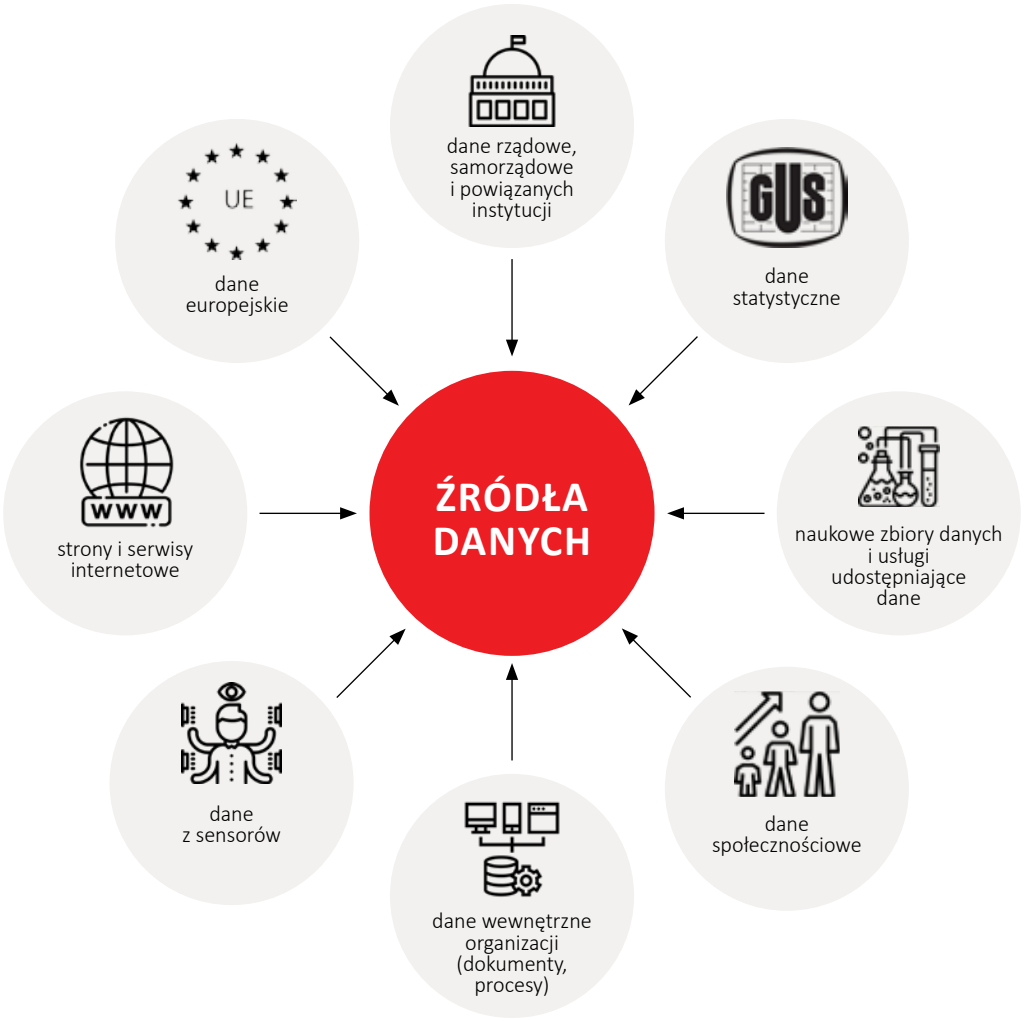
Mówiąc o podstawowych producentach (źródłach) zasobów danych, które będą opisane w dalszych częściach niniejszego raportu, postanowiliśmy podzielić je głównie ze względu na specyfikę pochodzenia i poziomu odpowiedzialności za publikowane dane. Korzystając z tego podziału, zaproponowanego na rys. **2.3**, wyróżnić możemy dane:

- udostępniane bezpośrednio przez administracje rządową i samorządową (głównie dane publiczne opisane w rozdziale **3**), również dane udostępnione w kontekście Unii Europejskiej,
- tematyczne; zbierane przez instytucje powiązane z administracją (np. opisane w sekcji **3.2.1**),
- zbierane i udostępniane przez Główny Urząd Statystyczny (sekcja **3.4.1**),
- otwarte; zbiory danych i usługi (bazy, interfejsy) publikowane na różnego typu licencjach o różnym poziomie otwartości – w szczególności zbiory o charakterze naukowym, pozwalające na budowanie unikalnych usług (np. opisane w sekcji **4.5** zbiory do głębokiego uczenia),
- dynamiczne usługi udostępniające dane z wykorzystaniem sensorów, często w infrastrukturze interfejsu sieciowego API[6],
- dane widoczne w serwisach i usługach dostępnych w Internecie[7]: same treści udostępniane w witrynach, wynikające bezpośrednio z działalności użytkowników (np. w tworzeniu usług crowdsourcingowych) oraz pośrednio (social media i komentarze),
- dane o charakterze unikalnym dla organizacji – wynikające z wewnętrznych procesów, logów wykorzystywanych systemów informatycznych (odwiedzin stron internetowych, wykorzystania usług, stanu dokumentów wewnętrznych, procesów biznesowych w organizacji). Nie są one bezpośrednim przedmiotem niniejszego raportu, ale są one punktem wyjściowym do jakiegokolwiek kreowania wartości dodanej, czy też innowacyjnych rozwiązań, ponieważ to właśnie wykorzystanie tych danych w połączeniu z przedstawionymi w raporcie może kreować przewagę konkurencyjną dla przedsiębiorstw i pozwalać na zwiększanie jakości usług administracji.

6 Application Programming Interface – interfejs programowania aplikacji, pozwalający na komunikację z udostępnioną aplikacją zarządzającą czujnikiem lub agregującą dane z nich pochodzące.

7 Opisane praktycznie w całości tego raportu, w zależności od ich tematyki w różnych rozdziałach.

RYSUNEK 2.3. **POTENCJALNE ŹRÓDŁA DANYCH MOŻLIWE DO WYKORZYSTANIA PRZEZ ADMINISTRACJĘ I PRZEDSIĘBIORCÓW**



Źródło: opracowanie własne.

Na podstawie tego podziału można również scharakteryzować udostępnione dane, określając, czego one dotyczą i wyróżniając różne typy danych[8]. Przykład tak wyróżnionych typów:

- **Źródło** – wpływa na wiarygodność danych; z wyróżnionych wcześniej: dane rządowe, samorządowe, instytucji rządowych, statystyczne, zbiory naukowe, dane z usług komercyjnych, dane z sensorów, crowdsourcing, dane wyekstrahowane: społecznościowe, ze stron i serwisów internetowych.
- **Pochodzenie** – wewnętrzne/zewnętrzne, istotne w celu odróżnienia zasobów unikalnych dla podmiotu. Oczywiście dokument urzędowy dla urzędu będzie daną wewnętrzną, jednocześnie jednak będzie mógł mieć status danych publicznych i otwartych – patrz sekcja **2.2**.
- **Aspekt techniczny** – opisuje strukturę i format danych, np. pojedynczy rekord, ich zbiór, plik, usługa (np. API). Powiązać go można z klasyfikacją open data przedstawioną w sekcji **2.2**, a także z formatami danych wykorzystywanymi do ich reprezentacji.
- **Typ danych** – tekst, liczba, obraz, dźwięk, graf, sekwencja, proces; głównie w celu określenia metod możliwych do stosowania dla analizy tych danych.
- **Domena** – wskazującą na tematykę (często naukową) tych danych lub kategorię, np. akty prawne, dane społecznościowe. Każdy przykład danej może mieć przypisane wiele takich kategorii tematycznych.

Powyższe charakterystyki mogą zostać poszerzone o dodatkowe wymiary, np. czas czy geografię. Dane geograficzne mogą pochodzić zarówno ze źródeł rządowych (patrz sekcja **3.5**), jak i z usług komercyjnych, takich jak Google Maps, czy usług zbudowanych w oparciu o inicjatywy crowdsourcingowe, jak OpenStreetMap opisane w sekcji **4.6.1**.

Dane mogą być również udostępnione na różnym poziomie **agregacji** ze względu na wymogi bezpieczeństwa, anonimizacji lub reprezentatywność próbki. Mówiąc o poziomach agregacji możemy wyróżnić:

- dane na poziomie krajowym,
- dane zagregowane na poziomie jednostek geograficznych (województw, powiatów, gmin),
- dane udostępniane na poziomie mniejszym niż gmina, dotyczące dzielnic lub innych arbitralnych jednostek podziału,
- dane o dostępie bezpośrednim (głównie sensory) – udostępniane na poziomie jednostki zbierającej. Oznaczają one możliwość dostępu do danych na poziomie danego obiektu (oddziału jednostki gospodarczej), a także pojedynczego czujnika.

8 Przyjęte tutaj określenia mają charakter typowo użytkowy.

Przykładowo, opisując dane z perspektywy jednego z urzędów rządowych przetwarzającego elektroniczne dokumenty można wskazać różne typy danych zaprezentowane w tabeli **2.3**. Wskazując na **dane najbardziej istotne z perspektywy kreowania innowacyjności i przewagi konkurencyjnej** należy wskazać dane o następujących cechach:

- niski poziom agregacji,
- duży wolumen lub wskazanie na reprezentatywną próbę,
- udostępnione w zautomatyzowany sposób,
- zgodne ze standardami open data (patrz sekcja **2.2**).

Przykłady te są dość ogólne, ale pozwalają na klasyfikację danych według zysku z ich wykorzystania. Korzystanie ze źródeł danych o charakterystykach wymienionych wyżej pozwala na budowanie użytecznych dynamicznych usług, np. na analizę natężenia dźwięku, ruchu i jakości powietrza (sensory), badania stanu techniki (zbiory publikacji naukowych) i wiele innych. W poniższym raporcie przy zakończeniu każdego z rozdziałów zostaną krótko scharakteryzowane potencjalne możliwości wykorzystania zaprezentowanych zbiorów.

TABELA 2.3. **MOŻLIWE PRZYKŁADY IDENTYFIKACJI DANYCH**

ŹRÓDŁO (PERSPEKTYWA RAPORTU)	POZIOM TECHNICZNY	POCHODZENIE	TYP	DOMENA	DODATKOWE WYMIARY	POZIOM AGREGACJI
DOKUMENT ELEKTRONICZNY W FORMACIE PDF NALEŻĄCY DO URZĘDU						
administracja rządowa / samorządowa	plik (pdf)	wewnętrzne	tekst	Dokumentacja elektroniczna, zarządzanie, administracja	geografia (miasto), osoba wydająca	brak
DOKUMENT ELEKTRONICZNY W FORMACIE PDF NALEŻĄCY DO URZĘDU						
administracja rządowa / samorządowa	plik (pdf)	wewnętrzne	tekst	Dokumentacja elektroniczna, zarządzanie, administracja	geografia (miasto), osoba wydająca	brak
REKORD W BAZIE DANYCH OPISUJĄCY STAN DOKUMENTU						
	rekord relacyjnej bazy danych	wewnętrzne	tekst i liczby (opis obiektu)	dokumentacja elektroniczna, zarządzanie, administracja, bazy danych		brak
LICZBA LUDNOŚCI W POWIECIE						
Źródło statystyczne (GUS)	liczba	zewnętrzne (GUS)	liczba, dane statystyczne	statystyka, demografia		agregacja geograficzna na poziomie powiatu

ŹRÓDŁO (PERSPEKTYWA RAPORTU)	POZIOM TECHNICZNY	POCHODZENIE	TYP	DOMENA	DODATKOWE WYMIARY	POZIOM AGREGACJI
STATYSTYKI PRZESTĘPSTW DLA DZIELNICY MIASTA DLA 2017 ROKU W PODZIALE NA MIESIĄCE						
Źródło statystyczne (KGP)	zbiór rekordów	zewnętrzne (KGP)	liczba, dane statystyczne	statystyka, przestępczość	kategorie przestępstw	agregacja geograficzna na poziomie dzielnicy, czasowa na poziomie miesięcy
USŁUGA UDOSTĘPNIAJĄCA ZANIECZYSZCZENIE POWIETRZA						
instytucja rządowa	usługa (API), udostepnia- jaca zbiór rekordów (JSON)	zewnętrzne (IMGW)	tekst i liczby	ochrona środowiska, przetwarzanie danych z sensorów		brak
NATĘŻENIE RUCHU NA NA ULICY DLA POJEDYNCZEGO SENSORA						
administracja rządowa / samorządowa	usługa (API), udostepnia- jaca zbiór rekordów (JSON)	zewnętrzne (administracja lokalna)	liczba, dane z sensorów	transport, przetwarzanie danych z sensorów		brak
administracja rządowa / samorządowa	zbiór rekordów	wewnętrzne	tekst, liczby, czas, sekwencja	procesy biznesowe, zarządzanie, administracja,		brak
PRZEBIEG PROCESU URZĘDOWEGO						
administracja rządowa / samorządowa	zbiór rekordów	wewnętrzne	tekst, liczby, czas, sekwencja	procesy biznesowe, zarządzanie, administracja,		brak
ZBIÓR DANYCH Z ZAKRESU GENETYKI						
naukowe zbiory danych	zbiór rekordów (CSV)	zewnętrzne	tekst, liczby...	genetyka		-
TREŚĆ ARTYKUŁU NA STRONIE INTERNETOWEJ PORTALU DZIENNIKARSKIEGO						
strona / serwis internetowy	tekst/ hipertekst (HTML)	zewnętrzne	tekst, strona www	dziennikarstwo, analiza tekstu (NLP), przetwarzanie informacji (Information Retrieval)	odnośniki do innych artykułów, popularność	brak
ZDJĘCIE UMIESZCZONE W SERWISIE INTERNETOWYM FLICKR.COM						
dane połecznościowe	obraz (jpg)	zewnętrzne	obraz	analiza obrazu, Social Media		brak
ZBIÓR WIADOMOŚCI NA TWITTERZE DOTYCZĄCYCH TAGU #MPIT						
dane połecznościowe	zbiór rekordów zawierających tekst	zewnętrzne	tekst, social media	analiza tekstu (NLP), Social Media		na poziomie tagu (z dostępem do pojedynczych rekordów

ŹRÓDŁO (PERSPEKTYWA RAPORTU)	POZIOM TECHNICZNY	POCHODZENIE	TYP	DOMENA	DODATKOWE WYMIARY	POZIOM AGREGACJI
NAGRANE POSIEDZENIE Z WYSTĄPIENIA SEJMOWEGO						
administracja rządowa	plik (dźwiękowy mp3)	zewnętrzne	plik dźwiękowy	analiza dźwięku, rozpoznawanie mowy, AI		brak
GRAF LICZBY MAILI WYSŁANYCH POMIĘDZY PRACOWNIKAMI						
dane wewnętrzne organizacji	zbiór rekordów (graf)	wewnętrzne	liczby, graf	analiza grafów i połecznosciowa	czas	na poziomie pojedynczego pracownika
POZYCJA GROBU ZMARŁEJ OSOBY UTRZYMYWANA W USŁUDZE MIASTA POZNAŃ						
administracja samorządowa	usługa (API), udostępniająca zbiór rekordów (GeoJSON)	zewnętrzne	tekst i liczby (opis obiektu)	dane geograficzne, rejestry państwowe	geografia	brak

Źródło: opracowanie własne.

2.4. Nasycenie danymi

Podobnie jak w przypadku danych nasycenie danymi ma wiele znaczeń, przy czym rozbieżność definicji jest jeszcze większa ze względu na kontekst użycia. Wspólnym mianownikiem jest próba szacowania, czy już wszystko wiemy, czyli: czy nowe dane coś wnoszą do naszej analizy?

Bardzo szeroko o nasyceniu danymi pisze się w naukach społecznych, szczególnie w kontekście prowadzenia wywiadów. Teoretyczne nasycenie danymi jest pojęciem z zakresu badań jakościowych, stosowanym w teorii ugruntowanej (grounded theory[9]). Oznacza ono osiągnięcie takiego stanu przez badaczy, że dalsze zbieranie próbek nie będzie prowadziło do pozyskania nowych informacji w odniesieniu do pytania badawczego. Przejawia się to tym, że prowadząc kolejne wywiady, badacze uzyskują te same kategorie w odpowiedziach, instancje się powtarzają, zatem mogą być empirycznie przekonani, że ich kategorie zostały nasycone, opisy są dostatecznie bogate i można przystąpić do prac nad teorią [**155**]. Nasycenie czasami podnoszone jest do rangi zasady metodologicznej badań jakościowych. Osiągnięcie nasycenia wskazuje na podstawie zebranych danych, że dalsze zbieranie danych jest bezcelowe [**152**]. Saunders badał również rolę nasycenia danych w różnych metodykach badawczych.

Nasycenie danymi możemy również odnieść do analiz ilościowych. W rozumieniu uczenia maszynowego nasycenie danymi to taki stan algorytmu, w którym podanie większej ilości danych nie polepsza działania klasyfikatora. W kontekście uczenia szczególnie ważna jest możliwość douczania modeli. I tak, przyjmuje się, że klasyczne algorytmy ML szybciej osiągają nasycenie danymi, natomiast głębokie sieci neuronowe potrafią polepszać swoje działanie wraz z dostarczaniem dodatkowych danych.

9 Spójny system metod jakościowych opisany w książce The Discovery of Grounded Theory (1967) przez A. Straussa i B. Glasera.

W obu powyższych podejściach nowe dane stają się redundantne w stosunku do danych już zebranych. Trudno nie zauważyć przecięć z przesytem informacyjnym (information overload). Ten jest jednak częściej rozumiany jako niezdolność ludzkiego umysłu do przyjęcia kolejnych informacji.

Nieco bardziej techniczne rozumienie nasycania danymi, określane również jako „danyfikacja", jest wykorzystywane przez [**122**], który definiuje to zjawisko jako „coraz większą zdolność do zbierania, przechowywania i analizowania danych, poprzez proliferację zintegrowanej technologii czujników i łączności".

Coraz większa zdolność do zbierania danych jest głównym wyzwaniem dla tych, którzy odpowiadają za ich przechowanie i analizę. Takie rozumienie nasycenia danymi zbliża nas do docelowej interpretacji tego pojęcia w raporcie. Na przykład autorzy [**54**] piszą o nasyceniu danymi w systemach business intelligence (BI). Według nich jest ono spowodowane „silną tendencją do wprowadzania do systemów BI wszystkich możliwych danych bez względu na to, czy są potrzebne, jak długo są potrzebne, na jakim poziomie są potrzebne (detaliczne czy zagregowane)." Powoduje to dwojakie reperkusje: przeładowane systemy informacyjne działają coraz wolniej, a sami użytkownicy nie mogą łatwo znaleźć właściwych raportów. Zaleca się, aby już na poziomie analizy wymagań funkcjonalnych określić, co i do czego jest użytkownikom potrzebne. Podsumowując, nasycenie danymi przedstawiane jest jako jedno z zagrożeń dla wdrożeń BI.

Od interpretacji mikroekonomicznej przechodzimy do interpretacji makroekonomicznej nasycenia danymi. Nasycenie gospodarki czy danej branży danymi determinuje zdolność do stosowania inteligentnych algorytmów. Nasycenie danymi określa się na podstawie tego, jak intensywnie wykorzystywane są dane w celu monetyzacji. Raport [**164**] jako skrajne przykłady podaje dwie branże: bankowość, gdzie wykorzystanie dużych i różnorodnych zasobów danych stało się standardem, szczególnie, jeśli mówimy o fintechach; oraz budownictwo, gdzie głównym czynnikiem produkcji jest ludzka praca, a przetwarzane dane mają przede wszystkim postać analogową (np. dokumentacja, projekty). Ten sam raport podkreśla, że nasycenie danymi polskiej gospodarki jest stosunkowo niskie. Istotne jest spojrzenie wieloaspektowe, a więc nie tylko samo posiadanie (zbieranie) danych, ale również zarządzanie nimi, integracja czy udostępnianie.

Kwestia nasycenia danymi w raporcie poruszane są wielokrotnie, choć nie są tak bezpośrednio nazwane. Na przykład zagadnienie dostępności danych w poszczególnych branżach szczegółowo było badane przez Open Data Barometer (patrz sekcja **3.3.6**.).

Zatem najważniejszym elementem praktycznym z omawianego tematu danych jest wskazanie potencjału wynikającego z ich wykorzystania. Po zapoznaniu się z różnymi definicjami i klasyfikacjami danych w organizacji warto zastanowić się, jak organizacja może wykorzystywać dane.

W przypadku wykorzystania danych możemy mówić o:

- udostępnieniu ich bezpośrednio, bądź w przetworzonym stanie jako produkt,
- optymalizacji, poprawie jakości procesu, w tym automatyzacji części procesu decyzyjnego na podstawie wcześniej podejmowanych decyzji,
- analizie stanu obecnego,
- predykcji – oszacowaniu stanu przyszłego,
- personalizacji usług i profilowaniu, dostarczając usługi charakteryzujące się większym dopasowaniem do potrzeb klienta,
- wzbogacaniu usług o dodatkowe funkcjonalności,
- budowaniu nowych usług.

2.4.1. Proces „nasycania danymi" – budowanie wartości

Korzystanie z wielu, lub z wszystkich wyżej wymienionych zastosowań może świadczyć o dużej dojrzałości w pracy z danymi i wysokim poziomie cyfryzacji organizacji, który można utożsamić z poziomem **"nasycenia danymi" danej organizacji.** Jeżeli więc wiele organizacji w danej branży cechuje się wysokim stopniem wykorzystania danych (także z wielu źródeł), możemy mówić o branży czy sektorze o "wysokim nasyceniu danymi". Samo jednak wykorzystanie danych nie kreuje wartości samej w sobie. Jednokrotna analiza bądź zatrudnienie specjalistów bez spójnego działania organizacji i wdrożenia odpowiedniego procesu kreowania wartości nie pozwoli na skorzystanie z wysokiego poziomu nasycenia danymi w sposób efektywny.

Mówiąc o kreowaniu wartości z danych trzeba myśleć o **przetwarzaniu danych jako o procesie, a nie jednorazowej czynności**. Możemy w ten sposób postarać się określić ramy procesu nasycania danymi, którego główne etapy można odczytać z rysunku **2.4**. Biorąc za przykład jeden z modeli tworzenia łańcucha wartości z wykorzystaniem danych (z uwzględnieniem obszaru big data), można wyróżnić kilka etapów, które organizacja powinna uwzględnić w tym procesie.

Pierwszym etapem jest **zbieranie danych**. W pierwszym momencie nie wiemy dokładnie, jakie dane chcemy przetwarzać. Proces powinno się więc zacząć od zdefiniowania źródeł, formatu i sposobu dostępu do danych, które mają kreować dla organizacji wartość. Dodatkowo dane te mogą mieć charakter: a) multimodalny – gdzie mamy do czynienia z różnymi typami danych opisującymi ten sam proces lub obiekt, bądź b) multikanałowy – w którym mamy różne źródła danych opisujące ten sam proces w perspektywie wielu kanałów realizacji (np. użytkownicy mobilni i kanał tradycyjny).

Po uzyskaniu dostępu do danych (np. w postaci dynamicznie generowanego strumienia), potrzebne są odpowiednie narzędzia i kompetencje do ich **analizy**. Także w tym momencie dane już dostępne należy wzbogacić i połączyć z innymi dostępnymi źródłami. Należy zbadać różne sposoby wzbogacenia badanego źródła, np. o semantykę (znaczenie), poprzez ekstrakcję informacji z nietypowych źródeł (jak np. obrazy czy logi), a także połączenie z innymi źródłami, celem nadania analizie dodatkowego wymiaru.

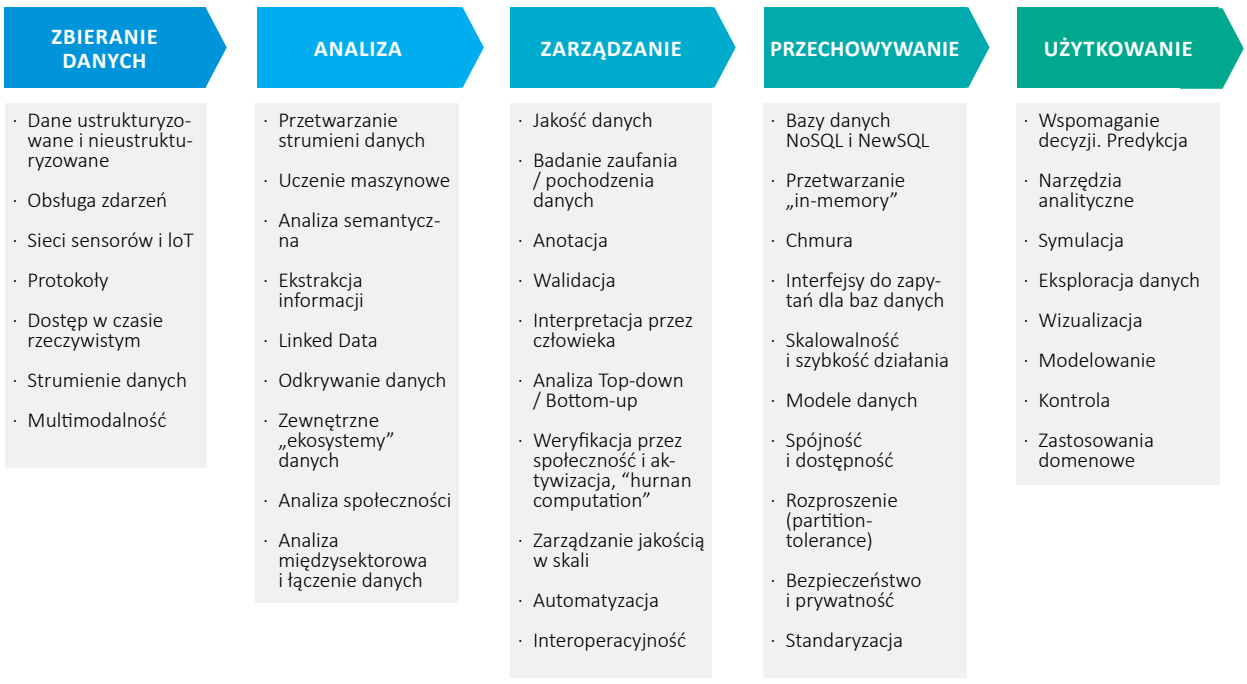
Zaprojektowanie analizy danych po ich uprzednim przetworzeniu jednak nie kreuje jeszcze wartości. Potrzebne jest **zarządzanie uzyskanymi wynikami** – badanie jakości i walidacja rezultatów. Niezbędne jest tutaj włączenie człowieka do oceny procesu analitycznego zaproponowanego przez algorytmy analityczne. To na tym etapie największą rolę odgrywa wspólny udział specjalistów-analityków i menadżerów dziedzinowych. Nie należy zbytnio ufać dobrym wynikom algorytmów, lecz starać się wyjaśnić ich działanie i konfrontować z rzeczywistością. Zarządzanie wynikami powinno być zaprojektowane w sposób, który umożliwia powrót do jednego z poprzednich etapów i zmianę źródła, parametrów, sposobu walidacji czy wykorzystanych zmiennych. Dopiero zapewnienie automatyzacji i interoperacyjności tego procesu pozwala na osiągnięcie wysokiego poziomu nasycenia pozwalającego na kreowanie odpowiedniego poziomu wartości. Analiza i zarządzanie nią powinno być procesem ciągłym i wydzielonym w działalności organizacji, opierającym się na czytelnych zasadach i dokumentacji tego procesu, z każdorazową walidacją wyników.

Dodatkowo przetworzone dane i uzyskane wyniki trzeba **przechowywać** w efektywny sposób, zachowując odpowiedni balans pomiędzy szybkością a bezpieczeństwem przechowywanych danych. Sposób przechowywania nie powinien wynikać z obecnej architektury firmowej, lecz z celowości przeprowadzanych analiz. Na tym etapie trzeba również uwzględnić tworzenie procesu analitycznego pozwalającego na odnoszenie się do poprzednich wyników analiz danych – same analizy powinny być dostępne w ten sam sposób jak dane, które posłużyły do ich budowy. Organizacja powinna pozwolić na odtworzenie poprzednich wyników, badając efektywność tego procesu. Umożliwi to samodzielne korzystanie z przygotowanych analiz na poziomie operacyjnym i przygotowywanie innowacyjnych analiz.

Korzystanie z ustandaryzowanego procesu w organizacji pozwala na efektywne użytkowanie danych do wielu wyżej wymienionych celów w sposób, który umożliwia kontrolę nad całością procesu i łatwe odnalezienie potencjalnych błędów lub wprowadzenie ulepszeń w procesie kreowania wartości. Można więc powiedzieć, że proces dobrego wykorzystania zasobów danych wymaga całej gamy narzędzi, kompetencji i kroków wspomagających, co przekłada się na nasycenie danymi danej organizacji.

W związku z tym, że proces ten zakłada automatyzację i parametryzację, szczególnie istotne są dla niego rzetelne i sprawdzone dane dostępne w sposób dynamiczny. Większość instytucji posiada przynajmniej jedno źródło takich danych, chociaż różnią się one znacznie w zależności od domeny (clickstreamy, dane sprzedażowe, produkcyjne itd.), jednak **każda organizacja może postarać się wzbogacić proces budowania wartości o wysokiej jakości dynamiczne zbiory "open data", czyli otwartych danych, dostępnych dla wszystkich podmiotów.** Ich powszechność i równy dostęp nie przekreśla przydatności i możliwości budowania przewagi konkurencyjnej z ich wykorzystaniem, ponieważ nie wszystkie podmioty potrafią wykorzystywać te dane, a dodatkowo mogą one tworzyć unikalną wartość dodaną w połączeniu z danymi, które przedsiębiorstwo już posiada. Mówiąc o zyskach z wykorzystania wyżej wymienionego procesu można wymienić m.in.: możliwość optymalizacji działania organizacji i większego wglądu w jej wewnętrzne procesy, możliwość automatycznego podejmowania decyzji, zdolność do oferowania spersonalizowanych usług wysokiej jakości. Jednak potencjalne korzyści mogą także wynikać z interoperacyjności i szybszej komunikacji z partnerami biznesowymi poprzez zautomatyzowane kanały komunikacyjne, a także tworzenie wirtualnych organizacji[10].

RYSUNEK 2.4. **PROCES "NASYCANIA DANYMI" – BUDOWANIA WARTOŚCI Z WYKORZYSTANIEM DANYCH DLA ORGANIZACJI**



Źródło: Grafika wykonana na podstawie [37]

10 https://mfiles.pl/pl/index.php/Organizacja_wirtualna.

2.4.2. „Nasycenie danymi" w różnych obszarach gospodarki

Szanse upatrywane w nasyceniu danymi podzielane są także przez Komisję Europejską, która definiuje pojęcie spójnego „europejskiego rynku cyfrowego" czy „rynku danych" [11]. Podejście to ma za zadanie zbudowanie silnej interoperacyjności i współpracy w obszarze zarządzania danymi i ich wykorzystania, zarówno pomiędzy administracją a przedsiębiorcami (A2B), jak i w modelu B2B. Jako że żadna organizacja nie działa w próżni i musi utrzymywać kontakt ze swoim otoczeniem gospodarczym, w tym partnerami biznesowymi i klientami, wysokie nasycenie danymi może pozwolić osiągnąć korzyści dużo większe niż tylko optymalizacja podstawowych procesów organizacji. Zauważając, że dane i ich wymiana (w szczególności mówiąc o sensorach i IoT) może zarówno wspomagać pracę pomiędzy firmami, ale także pozycjonować prywatnych właścicieli danych jako dostawców usług dla administracji rządowej w modelu usług B2A (ang. Business to Administration / Government), eksperci Komisji Europejskiej wskazują wiele zalet dla tych modeli.

Wspominając o procesie nasycania danymi, warto zwrócić uwagę w szczególności na sektory, które istotnie mogą skorzystać na wysokim poziomie kompetencji wykorzystywania danych w swoich organizacjach. Mówiąc o sektorach gospodarki, dla których szansa na skorzystanie z danych daje największe możliwości, można wymienić branże[12]:

- **Usługi finansowe** – ze szczególnym uwzględnieniem usług bankowych, branży fin-tech i usług ubezpieczeniowych. Monitorowanie transakcji, wykrywanie fraudów za pomocą nowych technik, np. biometria behawioralna [**86**], przetwarzanie języka naturalnego. Szansa pojawia się także w przetwarzaniu danych w postaci blockchain, czy smart-kontraktów w ubezpieczeniach skierowanych do klientów nieinstytucjonalnych.
- **Przemysł** – zgodnie z ideą przemysłu 4.0 i narastającego napływu danych z sensorów IoT [**136**] obszary produkcyjne będą skłonne do coraz większej automatyzacji procesów, a także do ciągłej ich optymalizacji dzięki algorytmom sztucznej inteligencji, wykrywającym błędy i pomagającym w predykcji poziomu produkcji.
- **E-commerce** – w obszarze sprzedaży internetowej przede wszystkim wyzwaniem będzie analiza wielu źródeł danych o kliencie, często z różnych urządzeń. Wykorzystanie danych to przede wszystkim integracja danych z wielu kanałów i umożliwienie dynamicznej personalizacji usług [**148**]. Możliwość analizy strumieni danych napływających z clickstreamów pozwoli na zbudowanie lepszych usług dla klientów i przyczyni się do zmniejszenia przewagi takich firm jak Amazon.

11 https://ec.europa.eu/digital-single-market/en/guidance-private-sectordata-sharing i https://ec.europa.eu/digital-single-market/en/big-datavalue-public-private-partnership

12 https://towardsdatascience.com/top-10-sectors-making-use-of-big-dataanalytics-be79d2301e79 i https://www.ibm.com/blogs/watson/2016/07/10industries-using-big-data-win-big/

- **Sportowa** – z uwzględnieniem usług „fitness" – korzystanie z IoT i sensorów podczas ćwiczeń fizycznych i sportu zarówno w wymiarze osobistym, jak i w skali sportu profesjonalnego, stworzy zapotrzebowanie na analityków i zacieśni współpracę pomiędzy uczelniami lub specjalistycznymi firmami mogącymi oferować usługi analityczne podmiotom związanym ze sportem.
- **Telekomunikacja i media** – z uwzględnieniem analizy Smart TV, multimodalnego i multikanałowego podejścia do analizy użytkownika, a także badania z wykorzystaniem zagregowanych danych do różnych celów[13], tj. analiza transportu drogowego[14] czy wykorzystanie danych na potrzeby statystyki publicznej [**46**].
- **Edukacja i szkolnictwo wyższe** – ze wspomaganiem kluczowych kompetencji potrzebnych dla ww. analizy. Skupienie się na bardziej praktycznych i warsztatowych formach nauki analizy danych, także z wykorzystaniem e-learningu. Rozwinięcie podstawowych kompetencji analitycznych, a także specjalizację w obszarach big data, data science i sztucznej inteligencji (AI), pozwoli na stworzenie kompetencji w zakresie przetwarzania danych w różnych sektorach gospodarki. Wymaga to jednak ścisłej współpracy uczelni z otoczeniem gospodarczym w kwestii dopasowywania programu i zmniejszenia sztywnych ram ograniczających programy studiów. Konieczna będzie także chęć współpracy ze strony przedsiębiorców, którzy muszą zacząć traktować uczelnie jako partnerów i outsourcing R&D, a nie tylko jako podmioty figurujące jako partnerzy grantowi.
- **Energetyka** – wraz z rozwojem nowych źródeł energii i procesu optymalizacji ich wykorzystania pojawiają się także liczne inicjatywy związane z sektorem smart-energy: pojęcia smart-grid i smart-meteringu [**85**, **175**]. Wyzwaniem będzie nie tylko budowanie sieci, ale także narzędzi do obsługi przez klientów detalicznych w ich domowych zastosowaniach.
- **Transport** – najbardziej związany z rozwojem AI i częściowej automatyzacji usług. Wdrożenia w tym sektorze w najbliższym czasie dotyczyć będą raczej optymalizacji procesu logistycznego we współpracy z firmami zajmującymi się handlem niż samochodów autonomicznych. Zastosowanie metod uczenia maszynowego i sztucznej inteligencji do optymalizacji procesów logistycznych będzie jednak wymagało bardzo specyficznego podejścia ze względu na dużą gamę algorytmów unikatowych dla tej dziedziny.
- **Usługi medyczne** – uspójnienie danych dla całego sektora medycznego wraz z korzystaniem przez odpowiednie jednostki badawcze z narzędzi analitycznych dla zanonimizowanych danych pozwoli nie tylko na lepsze zarządzanie infrastrukturą, ale i rozwój nowych technologii profilaktyki medycznej. Szansę, a także ogromne zagrożenie, można upatrywać w aplikacjach mobilnych skupionych wokół tematyki danych medycznych[15]. Twórcy takich aplikacji często swoje siedziby mają poza Europą, przez co mogą budować

13 Z przykładami zastosowań można się zapoznać czytając materiały pokonferencyjne z najbardziej znanej konferencji dotyczącej tematyki analizy danych telekomunikacyjnych – netmob.org.

14 https://www.dat.nl/en/solutions/mobility-patterns-frommobile-phone-data/

15 https://www.myheritage.pl/

swoje usługi analityczne na danych udostępnionych dobrowolnie przez użytkowników. Przy braku odpowiednich regulacji umożliwiających kontrolę i zachęcających do tworzenia takich usług na polskim gruncie, dane medyczne mogą być tym samym eksportowane do podmiotów poza kontrolą użytkowników, z dużą stratą dla możliwości rozwoju tego sektora.

- **Rząd i samorząd** – w rozumieniu cyfryzacji usług i udostępnianiu otwartych danych, których obszary będą poruszone w dalszej części raportu. Kluczowe tematy to rozwój cyfrowej dokumentacji i obiegu dokumentów we wnętrzu administracji, łączenie różnych zbiorów danych z nadawaniem im semantycznie spójnego znaczenia oraz budowa sieci sensorów IoT badających natężenie ruchu, zanieczyszczenia czy poziom hałasu w metropoliach.

Poza tymi obszarami można wskazać takie, dla których proces ten będzie się w późniejszym czasie, bądź wymaga to znaczącego nakładu środków. Takie obszary mogą pozwolić na budowę naprawdę innowacyjnego kapitału, ale charakteryzują się dużą niepewnością. Zaliczyć można do nich m.in. branżę budowniczą i jej wzbogacenie o modele BIM (ang. Building Information Modeling), technologie VR (ang. Virtual Reality) i AR (Augmented Reality).

2.5. Otwarte dane – przedmiot raportu

Przedsiębiorstwa dysponują danymi o różnorodnym charakterze, jednak ze względu na ich unikatowy charakter trudno jest zaproponować całościowe kompendium analizy każdego rodzaju danych dostępnego polskim firmom. W związku z powyższym raport ten będzie miał na celu określenie danych, które mogą pomóc wzbogacić i zbudować przewagę konkurencyjną dla wielu różnorodnych podmiotów i są dostępne na otwartym rynku danych. Uzasadnieniem podjęcia tematu otwartych danych w raporcie jest kilka przesłanek:

- Opisane tutaj przykłady otwartych danych mają zaznajomić czytelnika z danymi różnego typu i sposobami ich analiz. Są oczywiście pewne obszary, np. dane o procesach, logi, czy dane o zachowaniach klientów, które znajdują mniejsze odzwierciedlenie w wymienionym tutaj raporcie. Ich opis jednak przysparza trudności o tyle, o ile źródła takich danych są rzadko dostępne w odpowiedniej jakości, by przeprowadzić analizę wraz z uzyskaniem rzetelnych i weryfikowalnych wyników. Jednak fragmenty raportu dotyczące procesu analizy, źródeł danych, czy budowania modeli biznesowych powinny być dla tych organizacji wartością pozwalającą także na ulepszenie przetwarzania takiego rodzaju danych. Przegląd różnych źródeł i modeli przetwarzania zaproponowany w tym raporcie powinien także pozwolić na aplikację tych praktyk wewnątrz organizacji.
- Częściowym problemem pozostaje stan zasobów danych dostępnych w ogóle dla podmiotów gospodarczych i osób fizycznych. W celu określenia ich przydatności, zaznajomienia ze sposobami analizy i wskazania pewnych luk w ich implementacji, zostaną opisane najważniejsze wnioski z praktyk udostępniania otwartych danych.

- Zakres kompetencji analitycznych w obszarach: AI, data science czy big data jest spójny bez względu na to, czy mówimy o danych otwartych, czy wewnętrznych danych firmowych. Ze względu na duży wzrost zapotrzebowania na takiego rodzaju kompetencje, w niniejszym raporcie zostaną przedstawione kluczowe obszary pozwalające na przetestowanie takich kompetencji dla podmiotów chcących się podjąć rozwoju w tych obszarach. Jednocześnie więc wykształcenie kompetencji w obszarze open data pozwoli na ich łatwy transfer do biznesu.
- Otwarte dane [**79**], a także dane dostępne powszechnie w sposób dynamiczny z różnego rodzaju interfejsów sieciowych API [**8**], pozwalają na budowanie nowych, innowacyjnych usług, których przykłady także chcemy podać w tym raporcie. Usługi takie budowane są często z wykorzystaniem nowych metod analitycznych, we współpracy z sektorem nauki. Takie wspólne inicjatywy w zakresie korzystania z open data pomiędzy firmami, sektorem badań i rozwoju a administracją rządową mogą się przyczynić do znaczącego zwiększenia innowacyjności i rozwoju kompetencji potrzebnych do kreowania wartości z wykorzystaniem danych.

Korzyści z open data, zarówno dla administracji, jak i przedsiębiorstw wykorzystujących je do wzbogacenia swoich usług lub tworzenia całkowicie nowych, pozostają tematem bardzo szerokiej dyskusji w literaturze [**79**, **114**, **150**]. Wymienia się między innymi:

1. **Poziom rządowy:** wzrost transparentności działania, wzrost odpowiedzialności demokratycznej, zwiększenie zaufania do rządu i znaczące zwiększenie partycypacji obywateli w procesie decyzyjnym (zgodnie z ideą e-government), zmniejszenie wykluczenia cyfrowego i stygmatyzacji mniejszych podmiotów, poprawę istniejących i tworzenie nowych usług rządowych, a także zwiększenie satysfakcji ze świadczonych już usług. Dodatkowo, korzystanie z takich danych pozwala na polepszenie procesu legislacyjnego (z wykorzystaniem danych) oraz stymulowanie rozwoju wiedzy i kompetencji wśród obywateli i przedsiębiorców działających w kraju.
2. **Poziom gospodarki:** zwiększenie wzrostu gospodarczego i stymulację konkurencyjności wśród podmiotów gospodarczych, znaczącą stymulację procesu powstawania innowacji, zwiększenie inicjatyw do poprawy produktów i usług, kreowanie nowych sektorów i usług przynoszących zyski dla gospodarki, dużą dostępność i jakość informacji dla inwestorów i firm planujących inwestycje w kraju.
3. **Poziom operacyjny i techniczny poszczególnych organizacji:** zwiększenie powtórnego wykorzystania danych i zmniejszenie kosztów ich zbierania przez podmioty gospodarcze (poprzez przerzucenie części kosztów na sektor tworzący otwarte dane), zmniejszenie duplikacji danych (także przez inne instytucje publiczne) i zwiększenie ich jakości, optymalizację procesów administracyjnych, tworzenie oddolnych inicjatyw społecznych z wykorzystaniem danych, tworzenie wartości z łączenia danych.

Są to jedynie niektóre zalety wykorzystania otwartych danych, które stanowią przesłankę do poprawy sytuacji w obszarze danych dla wszystkich uczestników procesu – administracji rządowej i samorządowej, przedsiębiorców, instytucji publicznych, jak i konkretnych obywateli korzystających z danych w celach edukacyjnych, społecznych i gospodarczych. Dlatego więc duża część raportu poświęcona będzie otwartym zbiorom danych, jako inicjatywie służącej rozwojowi obszaru wykorzystania danych w przedsiębiorstwach.

Nie staramy się stawiać w roli ekspertów dla każdej działalności gospodarczej i każdego sektora. To przedsiębiorstwa wiedzą, w jaki sposób najlepiej skorzystać z danych będących w ich posiadaniu. W związku z tym w tym raporcie nie chcemy proponować przedsiębiorcom rozwiązań mających konkurować z ich wewnętrznymi systemami, nie chcemy też skupiać się na jednym obszarze rozwiązań gotowych, ponieważ mogą się one wtedy okazać nieinnowacyjne. Prezentując jednak możliwości wzbogacenia prowadzonej działalności o analizę nowych zbiorów, wykorzystując nowe metody i podejścia lub starając się korzystać z nowych modeli biznesowych, możemy zaproponować możliwości zwiększenia jakości i konkurencyjności świadczonych usług.

Prezentując różnego rodzaju dostępne zbiory, poza pokazaniem sposobu realizacji procesu kreowania wartości, możemy jednocześnie zaproponować organizacjom nowe źródła danych, mogące dla nich być postawą do tworzenia nowych usług.

Naszym zadaniem było:

- z pomocą ekspertów wskazać dostępne zbiory danych, które mogą być dla firm, administracji i pojedynczych programistów/analityków inspiracją do tworzenia nowych usług, a także użytecznym źródłem danych samym w sobie,
- zaprezentowanie maksymalnie szeroko sposobu wyszukiwania, analizy i budowania wartości z wykorzystaniem danych, który będzie jednakowy bez względu na to, czy mówimy o open data czy danych firmowych,
- zakreślenie znaczenia tematyki open data jako przykładu obszaru do kreowania wspólnej wartości dla wielu uczestników rynku. Tematyka otwartości i dostępności wydaje się w Polsce zaniedbywana, więc tym bardziej chcemy wskazać zalety rozwoju tego obszaru.

3 Dane publiczne – rząd dla społeczności

3.1. Wprowadzenie

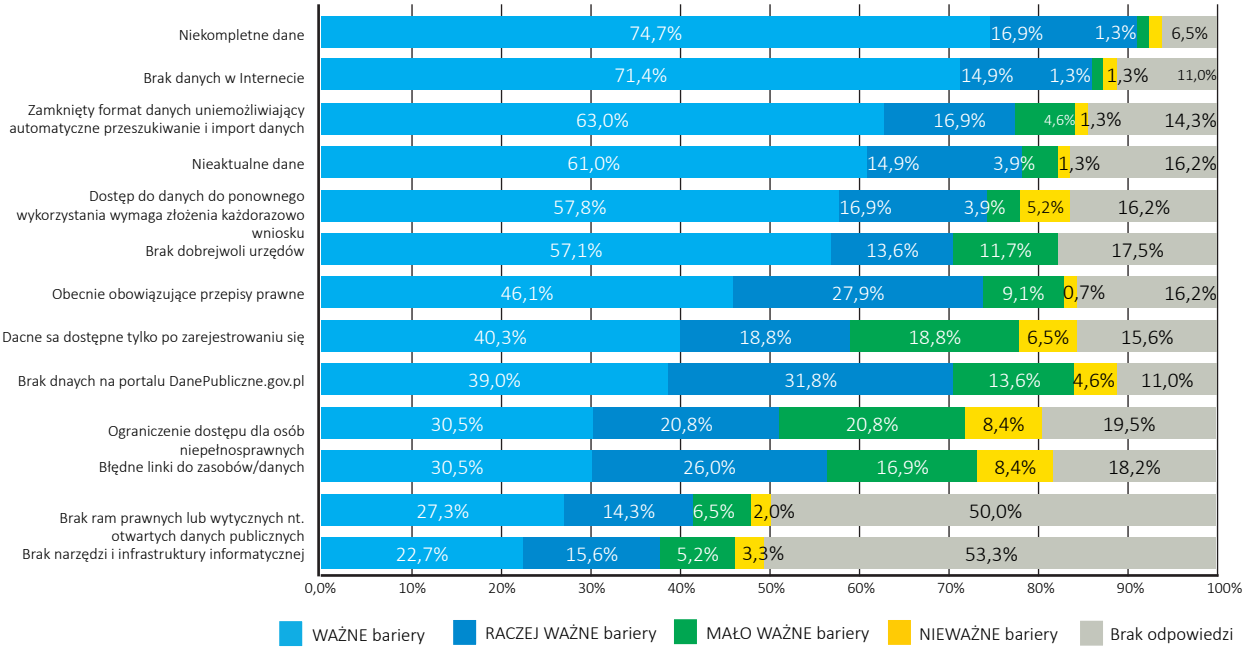
Niniejszy rozdział raportu ma za zadanie opisanie zasobów danych utrzymywanych przez administrację publiczną (rządową i samorządową), oraz organizacje działające w celu popularyzacji zbiorów danych dostępnych dla obywateli. Rozdział rozpoczyna się krótką charakterystyką aktualnych planów i strategii rządowych w zakresie udostępniania zasobów danych. Następnie krótko scharakteryzowane są zbiory rządowe i inicjatywy światowe oceniające stan zbiorów danych w Polsce i na świecie. Pozostała część raportu skupia się na opisaniu i scharakteryzowaniu w przystępny sposób najciekawszych zbiorów i usług zbudowanych z wykorzystaniem danych, wraz z charakterystyką najważniejszych obszarów rozwoju w zakresie open data.

Zgodnie z zaleceniami OECD w stosunku do otwartości i dostępności danych, w 2017 roku w Polsce ruszył projekt „Otwarte dane – dostęp, standard edukacja" w ramach poddziałania 2.3.1 Programu Operacyjnego Polska Cyfrowa. Jego celem jest zwiększenie poziomu otwartości, jakości i dostępności danych, które są w dyspozycji podmiotów publicznych [**119**]. W związku przygotowaniami do powyższego projektu, w wyniku przeprowadzonej ankiety (154 respondentów), zamieszczonej w kwietniu 2016 r. na stronach Ministerstwa Cyfryzacji, zidentyfikowane zostały bariery w dostępie do danych:

- niekompletne dane,
- dane nieudostępnione online,
- konieczność złożenia wniosku o dostęp do danych,
- zamknięty format danych,
- nieaktualne dane,
- brak spójnego katalogowania i wyszukiwania (np. brak danych na portalu danepubliczne.gov.pl, na co wskazała ponad 1/3 respondentów).

W ankiecie tej mogły wziąć udział osoby prywatne, a także: przedsiębiorcy, organizacje pozarządowe i środowiska akademickie. Wskazane bariery w dostępie do danych publicznych widoczne są również na rysunku **3.1**. Z badania wynika również, że aż 35% respondentów do pobierania danych publicznych używa API, czyli są to użytkownicy zaawansowani.

RYSUNEK 3.1. **ISTOTNOŚĆ BARIER W DOSTĘPIE DO DANYCH PUBLICZNYCH (KWIECIEŃ 2016)**



Źródło: wyniki ankiety z raportu Ministerstwa Cyfryzacji [118].

Ankietowani jasno wskazali, iż użyteczna byłaby możliwość korzystania z jednego centrum dostępu do danych publicznych. W nawiązaniu do powyższych ustaleń, od momentu analizy stanu obecnego w roku 2016 powstały inicjatywy mające na celu uruchomienie projektu poprawy stanu publicznych zasobów danych. Projekt uruchomiony został w czerwcu 2017. Określił on następujące bazy, które miały zostać otwarte lub przebudowane:

- KRS (Krajowy Rejestr Sądowy),
- BDL (Bank Danych Lokalnych),
- CEPiK (Centralna Ewidencja Pojazdów i Kierowców),
- BESTI@ (finanse jednostek samorządu terytorialnego),
- NFZ (Narodowy Fundusz Zdrowia) – statystyki nt. pacjentów, kolejki oczekiwania na świadczenia medyczne oraz Informator o umowach zawieranych przez NFZ z placówkami medycznymi.

Na podstawie powyższych działań mogą się pojawić pytania o przydatność udostępnionych zbiorów i możliwość wykorzystania także innych katalogów w celu wzbogacenia i poprawienia jakości świadczonych przez przedsiębiorców usług za pomocą danych. Zasadna może być także ocena obecnego stanu przygotowania świadczonych przez sektor publiczny usług zgodnych z paradygmatem otwartych danych i ich dojrzałości, ponieważ to one są podstawowym źródłem rzetelnych i weryfikowalnych danych w gospodarce. W celu m.in. oceny stanu i przygotowania do wykorzystania tych i innych dostępnych w gospodarce zbiorów powstaje niniejszy raport. Opisane zostaną zbiory nie tylko polskie, lecz także światowe inicjatywy, rankingi, serwisy i opracowania mające na celu wskazanie praktycznych aspektów użyteczności i dostępu do wyżej wymienionych danych.

3.2. Zakres danych publicznych

Mówiąc o globalnym ekosystemie udostępnionych, otwartych zbiorów można mówić o całym „rynku danych". Dane udostępniane przez administracje publiczną dotyczą głównie podstawowych informacji o gospodarce danego państwa i w przejrzysty sposób dają wgląd w inicjatywy legislacyjne i aktualny stan aktów prawnych obowiązujących w danym kraju. Starając się tematycznie uszczegółowić zakres publikowanych danych, można m.in. wyróżnić dane dotyczące:

- Obowiązujących aktów prawnych i pracy organów ustawodawczych.
- Pracy i dokumentacji związanej z pracami instytucji państwowych, publicznych i samorządowych.
- Danych statystycznych związanych ze wszystkimi działami gospodarki i administracji, a także ludności i zamieszkania.
- Danych dotyczących nauki, technologii, kultury i sportu.
- Rejestrów państwowych – patentów, rejestru pojazdów, rejestrów sądowych.
- Transportu, środowiska, zdrowia i profilaktyki zdrowotnej.

Niektóre z tych danych mogą być udostępnione jedynie w ograniczonym zakresie, zgodnie z możliwościami technicznymi (i potrzebą ich opracowania). Niektóre zbiory w związku z kwestiami bezpieczeństwa i zachowania prywatności mogą być udostępnione jako zanonimizowane lub jako zagregowany wynik.

Bardzo ważną kwestią, niezbędną do korzystania z towarów (tutaj zbiorów danych), jest możliwość ich oceny przy pomocy wybranych kryteriów. Jednym z podstawowych może być rzetelność i pewność poprawności udostępnionych danych. Jako że dane administracyjne są w związku z powyższym podstawą budowy ekosystemu open data, powstało wiele serwisów (w tym oficjalnie sygnowanych przez inicjatywy UE) poświęconych kompleksowej ocenie powyższych zbiorów.

3.2.1. Rejestry państwa

Administracja państwowa może (i zgodnie z wyżej wymienionymi standardami powinna) udostępniać szeroki zakres zbiorów danych. Celem niniejszego podrozdziału jest opisanie najbardziej rozbudowanych i uzupełniających się tematycznie serwisów i usług pozwalających na uzyskanie danych pochodzących z administracji publicznej.

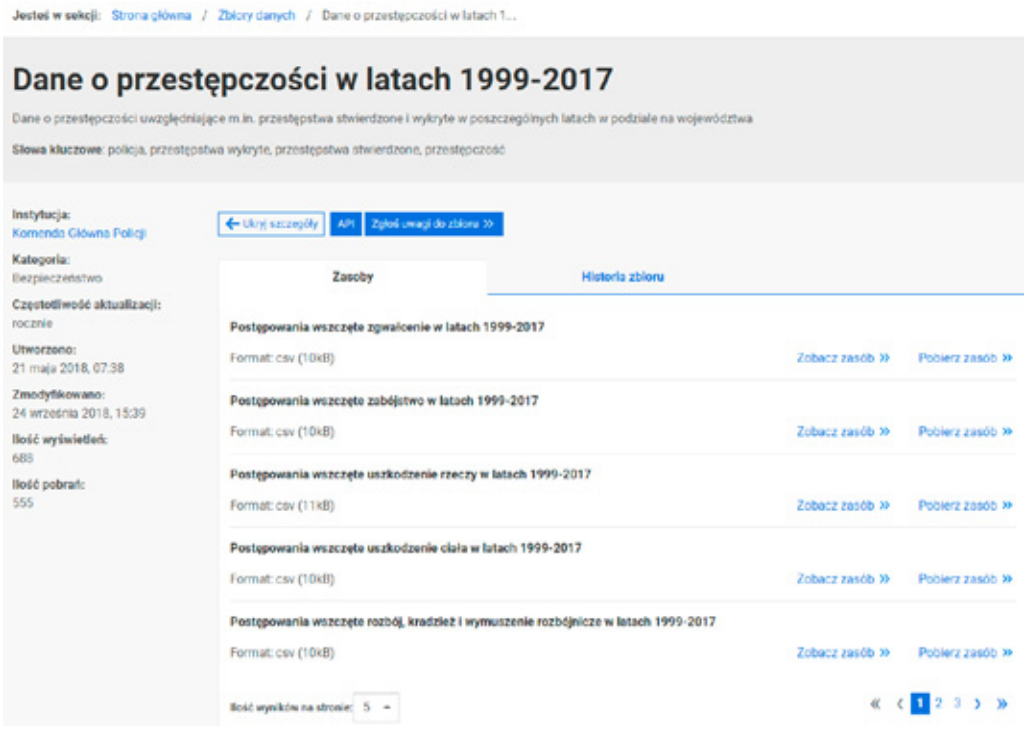
dane.gov.pl

W Polsce podstawowym źródłem danych publicznych jest serwis https://dane.gov.pl/. W styczniu 2019 r. udostępniał on ponad 11 tysięcy zbiorów z ponad 120 instytucji. Serwis pozwala na wyszukiwanie zbiorów zgodnie z interesującą użytkownika tematyką oraz, co szczególnie ważne, udostępnia API pozwalające na przetwarzanie udostępnionych zbiorów. Serwis ten oferuje różnorodne zbiory danych – w tym dane dotyczące m.in. przestępczości, rejestracji pojazdów, zdrowia, środowiska, sportu i turystyki, rolnictwa, gospodarki i administracji.

Przykładowo wskazać można zbiór danych „Dane o przestępczości" w latach 1999-2017[1]. Widok na dany zbiór zaprezentowany jest na rysunku **3.2**. Dane posiadają kategorię, datę utworzenia i modyfikacji (możliwość sprawdzenia historii zmian). Pod ten zbiór danych podpiętych jest wiele zasobów, w formacie CSV – jest to 3 poziom open data. Przetworzenie więc takich danych może być wykonane za pomocą dowolnego oprogramowania, a nie musi być wykonywane za pomocą np. programu Excel.

1 https://dane.gov.pl/dataset/1012

RYSUNEK 3.2. **PRZYKŁADOWY ZBIÓR DANYCH Z SERWISU DANE.GOV.PL – STATYSTYKI DOTYCZĄCE PRZESTĘPCZOŚCI, KTÓRYCH ŹRÓDŁEM JEST KGP**



Źródło: https://dane.gov.pl/, katalog danych nr 1012

Aby dostać się do tego zasobu, możemy skorzystać również z udostępnionego interfejsu programowania aplikacji (API). W przypadku tego zbioru adres dostępu do niego jest następujący – https://api.dane.gov.pl/datasets/1012. Wynik uzyskany dzięki interfejsowi programistycznemu pozwala na zwrócenie informacji o zbiorze w sposób przetwarzalny automatycznie. Wynik uzyskany w ten sposób zaprezentowany jest w listingu **3.1**. Przedstawia on dane zwrócone w formacie JSON[2].

2 Jest to bardzo popularny format opisu danych, w szczególności w przypadku danych dostępnych w interfejsach API. Podstawowe informacje o nim można uzyskać chociażby w języku polskim pod adresem https://pl.wikipedia.org/wiki/JSON lub korzystając z dokumentacji formatu http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf.

KOD ŹRÓDŁOWY 3.1. **PRZYKŁAD WYNIKU ZWRÓCONEGO ZA POMOCĄ API SERWISU DANE.GOV.PL DLA ZBIORU 1012**


```
{
    "data": {
        "type": "dataset",
        "attributes": {
            "views_count": 691,
            "url": "",
            "license_name": "other-pd",
            "notes": "<p>Dane o prz\u0119stepczo\u015bci uwzgledniaj\u0105ce m.in.
                ↳ przest\u0119pstwa stwierdzone i wykryte w poszczeg\u00f3lnych
                ↳ latach w podziale na wojew\u00f3dztwa</p>\n",
            "followed": false,
            "update_frequency": "rocznie",
            "category": {"id": 136, "title": "Bezpiecze\u0144stwo"},
            "license_condition_db_or_copyrighted": "",
            "title": "Dane o prz\u0119stepczo\u015bci w latach 1999-2017",
            "license_condition_modification": null,
            "license_condition_timestamp": null,
            "downloads_count": 556,
            "license_condition_source": true,
            "license_condition_original": null,
            "created": "2018-05-21 05:38:46.470342+00:00",
            "openness_scores": [1, 2, 3],
            "formats": ["pdf", "xlsx", "docx", "csv"],
            "tags": [
                "policja",
                "przest\u0119pstwa wykryte",
                "przest\u0119pstwa stwierdzone",
                "przest\u0119pczo\u015b\u0107"
            ],
            "license_condition_responsibilities": "",
            "modified": "2018-09-24 13:39:33.319025+00:00",
            "license_description": "Other (Public Domain)"
        },
        "relationships": [],
        "id": 1012,
        "links": {"self": "/datasets/1012"}
    },
    "included": [],
    "meta": {
        "language": "pl",
        "params": {},
        "path": "/datasets/1012",
        "rel_uri": "/datasets/1012"
    }
}
```

W zwróconym przez API wyniku widać chociażby unikalny identyfikator zbioru (wartość przy parametrze "id"), powiązaną z nim datę powstania ("created") i inne informacje, które dostępne są w serwisie www.gov.pl, pokazanym na rysunku **3.2**. Ten jeden katalog zawiera referencje do wielu zbiorów danych zawartych w liście "included". Lista ta zawiera odnośniki do poszczególnych zasobów (resources) dostępnych w tym zbiorze. Każdy taki zbiór również posiada swoje atrybuty, co zobaczyć można w listingu **3.2**. Najważniejszym parametrem w przypadku zbiorów udostępnionych w plikach jest więc "file_url" który zawiera bezpośredni odnośnik do zbioru i pozwala na jego automatyczne pobranie. Dzięki temu można uzyskać bezpośredni dostęp do pliku, którego znalezienie lub uaktualnienie normalnie wymagałoby odwiedzenia strony internetowej. Poprzez udostępnienie API i dzięki dodatkowym informacjom (np. o dacie modyfikacji, wielkości pliku) możliwa jest weryfikacja w sposób automatyczny. Pobrany wcześniej zbiór można również uaktualnić w sposób automatyczny, jeżeli pojawiła się jego nowsza wersja.

KOD ŹRÓDŁOWY 3.2. **PRZYKŁAD WYNIKU ZWRÓCONEGO ZA POMOCĄ API SERWISU DANE.GOV.PL DLA JEDNEGO Z ZASOBÓW W ZBIORZE 1012 – FRAGMENT LISTY "INCLUDED"**

```
[
    {
        "type": "institution",
        "id": 27,
        "attributes": {
                "title": "Komenda Główna Policji",
                "links": {"self": "/institutions/27"}
        },
    },
    {
        "type":"resource",
        "attributes":{
            "format":"csv",
            "downloads_count":9,
            "description":"<p></p>\n",
            "created":"2018-05-21 08:15:50.849171+00:00",
            "openness_score":3,
            "type":"file",
            "title":"Postepowania wszczete kradziez z włamaniem w latach 1999-2017",
            "file_size":11165,
            "views_count":32,
            "file_url":"https://www.dane.gov.pl/media/resources
                ↪ /20180521/postepowaniawszczete-kradziezzwlamaniem.csv",
            "modified":"2018-05-21 08:15:50.849171+00:00",
            "link":"https://www.dane.gov.pl/media/resources/20180521/
                ↪ postepowaniawszczete-kradziezzwlamaniem.csv"
        },
        "id":10426,
        "links":{"self":"/resources/10426"}
    }
]
```

Inne usługi państwowe

Wśród innych usług oficjalnie udostępnianych przez administrację wyróżnić można m.in.:

- **Dane meteorologiczne API IMGW** https://danepubliczne.imgw.pl/ – serwis udostępniający API prowadzony przez Instytut Meteorologii i Gospodarki Wodnej Państwowy Instytut Badawczy. Serwis ten udostępnia dane dotyczące m.in. temperatury, prędkości i kierunku wiatru, ciśnienia i opadów dla stacji meteorologicznych w Polsce oraz informacje o ostrzeżeniach pogodowych. Jest to bardzo dobry przykład interfejsu API realizującego w dobry sposób potrzebną usługę. Przykład odpowiedzi interfejsu w formacie JSON widoczny jest w listingu **3.3**. Poza danymi dostępnymi dynamicznie możliwe jest również pobranie agregatów przedstawiających kształtowanie się wyżej wymienionych danych w czasie.
- **Dane o jakości powietrza GIOŚ** http://powietrze.gios.gov.pl/pjp/content/api – podobną tematycznie usługą jest prowadzone przez Główny Inspektorat Ochrony Środowiska API wskazujące zbiory dotyczące jakości powietrza dla poszczególnych stacji pomiarowych. Można w ten sposób, analogicznie do powyższego przykładu, uzyskać listę wszystkich stacji pomiarowych[3]. Następnie wybierając daną stację (np. 729) uzyskać wszystkie zamontowane w niej sensory[4] – wynik widoczny jest w kodzie **3.4**. Następnie wybierając konkretny sensor, np. dotyczący pyłu PM10 dla tej stacji (id 4681), możemy sprawdzić jego aktualny poziom. Przykład takiego wyniku zobaczyć można w listingu **3.5**, gdzie pokazane są dane godzinowe.
- **Polska Bibliografia Naukowa i POLON** https://pbn-ms.opi.org.pl – Polska Bibliografia Naukowa jest rozwiązaniem prowadzonym pod egidą Ministerstwa Nauki i Szkolnictwa Wyższego, pozwalającym na sprawdzenie aktualnego stanu polskiej nauki. Możliwe jest dzięki niemu wyszukiwanie publikacji naukowych polskich autorów i jednostek naukowych. Moduł statystyczny pozwala na przeglądanie osiągnięć raportowanych do systemu POLON, a jego interfejs można zobaczyć na rys. **3.3**. Serwis, mimo iż posiada w tej chwili API, nie udostępnia go w otwarty sposób. Ze względu jednak na dostępny moduł wyszukiwania[5] system ten pozostaje praktycznym narzędziem mogącym służyć współpracy pomiędzy biznesem a jednostkami naukowymi. Dzięki niemu można m.in. znaleźć naukowców zajmujących się tematyką badań podejmowaną w przedsiębiorstwie. Ograniczenie polskich jednostek i naukowców działa w tym przypadku na korzyść, pozwalając na zawężenie kryterium wyszukiwania w stosunku do usług pozwalających na podobne wyszukiwanie[6], opisanych w rozdziale **7**.

3 http://api:gios:gov:pl/pjp-api/rest/station/findAll
4 http://api.gios.gov.pl/pjp-api/rest/station/sensors/729
5 https://pbn-ms.opi.org.pl/pbn-report-web/pages/search/
6 Jak http://scholar.google.com czy https://www.elsevier.com/search.

- **REGON** https://wyszukiwarkaregon.stat.gov.pl/appBIR/index.aspx oraz nowszy serwis https://api.stat.gov.pl/Home/RegonApi – są to portale pozwalające na dostęp do bazy REGON. Pierwszy z nich to standardowa wyszukiwarka nie pozwalająca na automatyczny dostęp (przykład przedstawiający Uniwersytet Ekonomiczny Poznaniu pokazany na rys. **3.4**.). Baza ta pozwala na uzyskanie wielu informacji o osobach prawnych, klasyfikacji ich działalności i danych kontaktowych. Druga z usług pozwala na dostęp do tego systemu poprzez interfejs API dla podmiotów komercyjnych i administracji publicznej.
- **CEIDG** https://prod.ceidg.gov.pl/CEIDG/CEIDG.Public.UI/Search.aspx – usługa prowadzona przez Ministerstwo Przedsiębiorczości i Technologii pozwala na wyszukiwanie informacji z rejestru CEIDG (wymaga wpisania kodu Captcha, co nie pozwala na automatyczne przetwarzanie wpisów). Jednak możliwe jest napisanie własnej usługi tworząc konto w serwisie CEIDG Data Store[7].
- **KRS** https://ekrs.ms.gov.pl/web/wyszukiwarka-krs – dostęp do systemu jest zapewniony poprzez wyszukiwarkę, co utrudnia odpytywanie automatyczne. Przeglądanie wymaga wysłania formularza. Po wpisaniu numeru KRS można przeglądać dokumenty finansowe powiązane z podmiotem[8].
- **Dane statystyczne GUS** https://api.stat.gov.pl/ i https://bdl.stat.gov.pl/bdl/start – są to bardzo szerokie tematycznie serwisy, które prowadzone są przez GUS i udostępniają dane statystyczne. Ze względu na ich zakres, opisane będą dalej w osobnym podrozdziale **3.4.1**.
- **CEPIK** https://historiapojazdu.gov.pl/ i bezpiecznyautobus.gov.pl – serwisy „historia pojazdu" i „bezpieczny autobus" udostępniają dane dotyczące pojazdów z systemu CEPIK. uzyskanie informacji wymaga podania numeru rejestracyjnego, numeru VIN i daty pierwszej rejestracji pojazdu. System ten jest bazą wielu komercyjnych usług i planowana jest jego rozbudowa[9] oraz poszerzenie zakresu możliwości automatycznego przetwarzania.

KOD ŹRÓDŁOWY 3.3. **PRZYKŁAD WYNIKU API IMGW POKAZUJĄCEGO STAN ODCZYTÓW SENSORÓW DLA STACJI JELENIA GÓRA**

```
{"id_stacji":"12500",
"stacja":"Jelenia G\u00f3ra",
"data_pomiaru":"2019-01-18",
"godzina_pomiaru":"16",
"temperatura":"-2.2",
"predkosc_wiatru":"3",
"kierunek_wiatru":"300",
"wilgotnosc_wzgledna":"64.8",
"suma_opadu":"0000000000.4",
"cisnienie":"1021.3"}
```

7 https://datastore.ceidg.gov.pl/; informacje można również otrzymać w referencyjnym linku dostępnym w serwisie dane.gov.pl https://www.data.gov.pl/dataset/646/resource/979.

8 https://ekrs.ms.gov.pl/rdf/pd/search_df

9 https://bazakonkurencyjnosci.funduszeeuropejskie.gov.pl/ publication/view/1129306

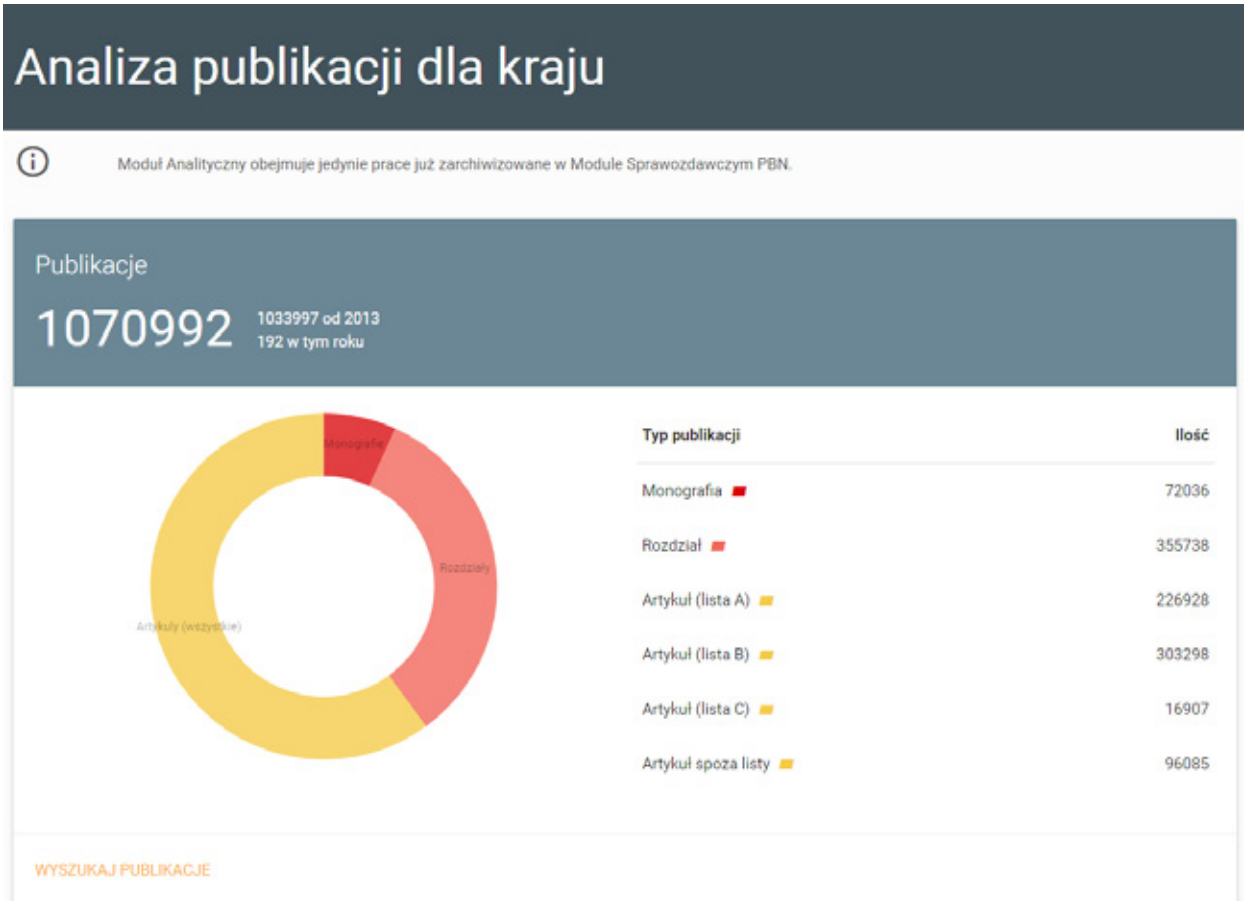
KOD ŹRÓDŁOWY 3.4. **PRZYKŁAD WYNIKU API IMGW POKAZUJĄCEGO WSZYSTKIE SENSORY DOSTĘPNE DLA STACJI 729**

```
[{"id":4676,"stationId":729,"param":
    {"paramName":"tlenek węgla",
    "paramFormula":"CO",
    "paramCode":"CO",
    "idParam":8}},

{"id":4679,"stationId":729,"param":
    {"paramName":"dwutlenek azotu",
    "paramFormula":"NO2",
    "paramCode":"NO2",
    "idParam":6}},

{"id":4681,"stationId":729,"param":
    {"paramName":"pył zawieszony PM10",
    "paramFormula":"PM10",
    "paramCode":"PM10",
    "idParam":3}},

{"id":4683,"stationId":729,"param":
    {"paramName":"dwutlenek siarki",
    "paramFormula":"SO2",
    "paramCode":"SO2",
    "idParam":1}}]
```

KOD ŹRÓDŁOWY 3.5. **PRZYKŁAD WYNIKU API IMGW POKAZUJĄCEGO WYNIKI POMIARU STĘŻENIA PYŁU PM10 DLA SENSORA O ID 4681**

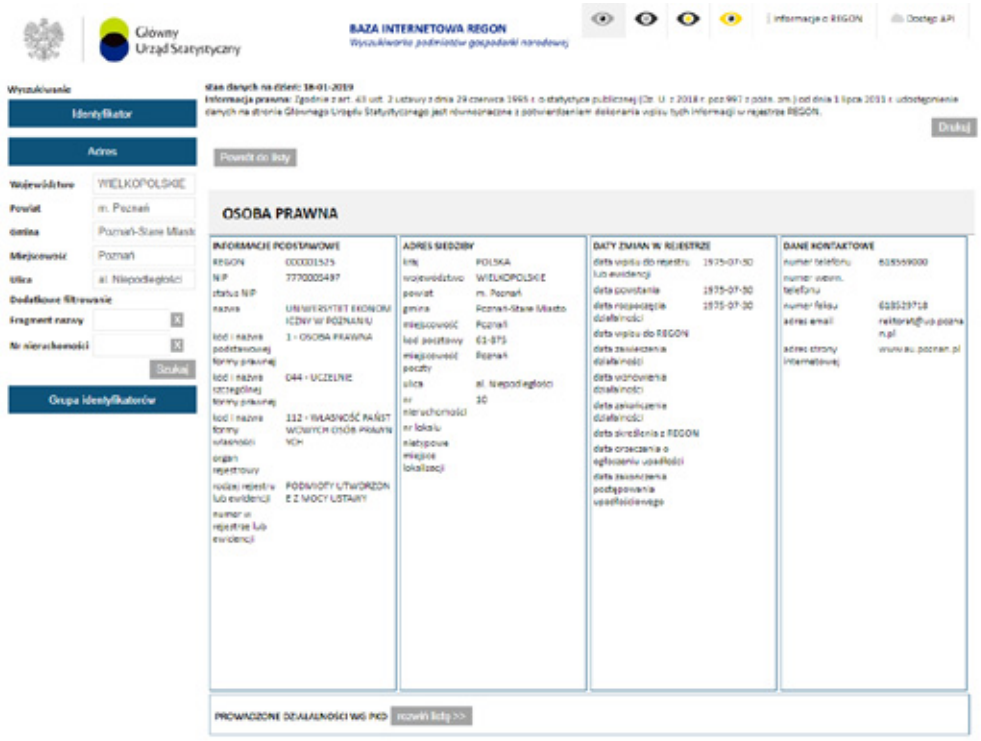
```
{"key":"PM10","values":[
    {"date":"2019-01-16 13:00:00","value":17.5444},
    {"date":"2019-01-16 12:00:00","value":43.6644},
    {"date":"2019-01-16 11:00:00","value":71.705},
    {"date":"2019-01-16 10:00:00","value":62.7922},
    {"date":"2019-01-16 09:00:00","value":67.4325},
    {"date":"2019-01-16 08:00:00","value":29.9222},
    {"date":"2019-01-16 07:00:00","value":16.3286},
    {"...": "..."},
    ]
}
```

RYSUNEK 3.3. **SERWIS POLSKA BIBLIOGRAFIA NAUKOWA (PBN) – STATYSTYKI DLA POLSKI**



Źródło: https://pbn-ms.opi.org.pl

RYSUNEK 3.4. **STRONA WYSZUKIWANIA DLA REJESTRU REGON – PRZYKŁAD WYSZUKIWANIA**



Źródło: https://wyszukiwarkaregon.stat.gov.pl/appBIR/index.aspx

Dodatkowo związane z powyższymi usługami są między innymi: **Baza wyroków Urzędu Ochrony Klienta i Konsumenta**[10], **Baza rejestru zamówień publicznych**[11] oraz **Krajowy Rejestr Dłużników niewypłacalnych**[12]. Są to usługi, które wraz z powyższymi są wykorzystywane do budowania serwisów udzielających informacji o podmiotach gospodarczych. Podsumowując, nie wszystkie usługi w tej chwili udostępniają API zgodne z najnowszymi standardami, przez co wymagane jest budowanie programów imitujących wyszukiwanie przez użytkownika (crawlerów), co stawia barierę wykorzystania dla małych podmiotów. Jednocześnie brakuje spójnego katalogu zawierającego powyższe usługi.

10 https://decyzje.uokik.gov.pl/bp/wyroki.nsf
11 https://ssdip.bip.gov.pl/zamowieniapubliczne/wyszukaj/advanced:1/
12 http://krdn.pl/sprawdzosobe

3.2.2. Usługi komercyjne i prowadzone przez fundacje

Na podstawie zbiorów, które zostały opublikowane przez instytucje państwowe, zbudowano wiele usług. Część z takich inicjatyw została opisana na portalu dane.gov.pl[13] w zakładce „aplikacje" lub podkreślona w różnorodnych raportach dotyczących otwartości danych w Polsce. Poniżej scharakteryzowane zostały przykłady usług, które można rozpatrywać jako zasoby do wykorzystania (choć pamiętać należy o tym, iż niektóre z nich mogą być odpłatne i udostępnione na warunkach określonych przez usługodawcę).

Inicjatywy fundacji, organizacji i osób prywatnych

Pierwsza kategoria usług to usługi prowadzone przez fundacje i osoby prywatne o charakterze wolnego dostępu.

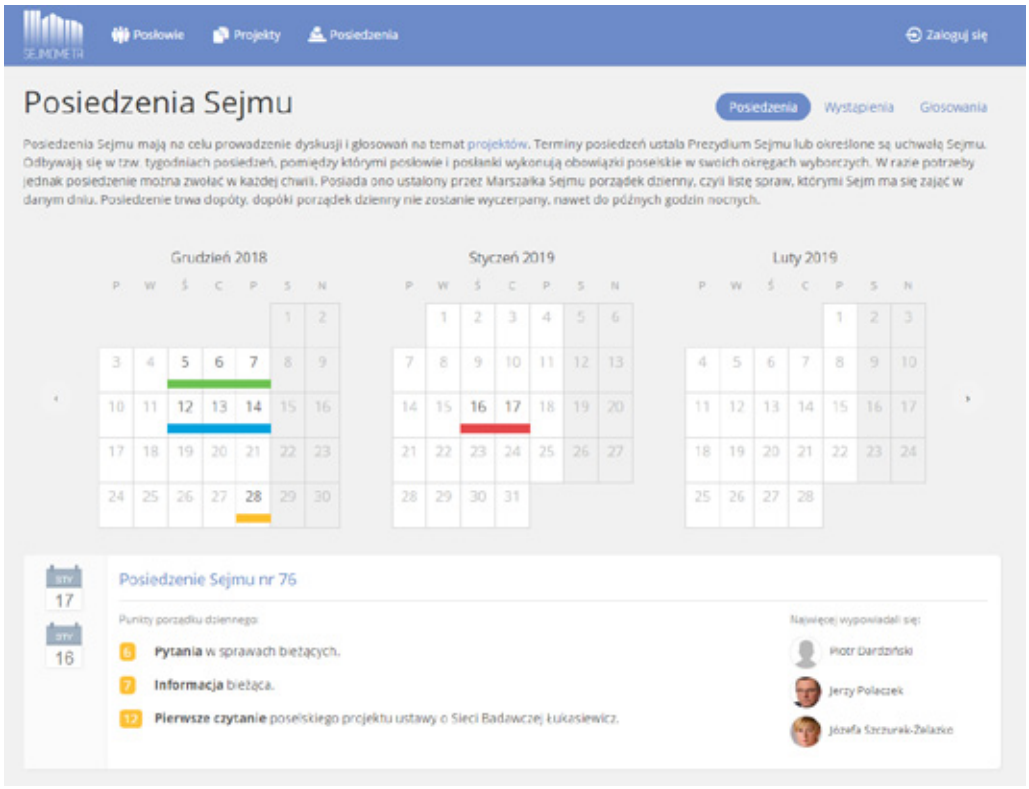
Przykładem takiego serwisu jest **mojepanstwo.pl**[14] – Fundacja ePaństwo utworzyła w roku 2016 witrynę umożliwiającą przeglądanie i korzystanie ze zbiorów danych udostępnianych przez polską administrację. Strona internetowa udostępnia dane o różnorodnej tematyce, od przetargów, raportów z systemu KRS czy działań administracji publicznej, samorządów oraz rejestru zamówień publicznych. W tym momencie serwis udostępnia 3 funkcjonalne API[15] służące do przeglądania: KRS, rejestru zamówień publicznych i tzw. sejmometr[16] pozwalający na przeglądanie m.in. treści wystąpień sejmowych. Poza tym serwis udostępnia następujące usługi:

- Zakładka NGO i usługa „Kto tu rządzi?" – pozwalające na przeglądanie funkcjonujących na terenie Polski organizacji pozarządowych, obejmujących: fundacje, stowarzyszenia, związki zawodowe i spółdzielnie. Natomiast usługa „Kto tu rządzi"[17] pozwala na przeglądanie podobnych informacji o strukturze organizacji rządowych i samorządowych. Serwis jest ciągle rozbudowywany, ale już w tym momencie pozwala m.in. na przeglądanie listy zbiórek publicznych.
- Podatki – serwis informacyjny pokazujący, jakie kwoty z wynagrodzenia konkretnej osoby statystycznie zasilają budżet państwa i w jakim stopniu są spożytkowane w budżecie.
- Środowisko / kultura – serwis prezentuje opracowania raportów dotyczących kultury (oparte w większości na badaniach GUS) oraz wykorzystuje wymienione w poprzednim rozdziale dane o jakości powietrza.

13 https://dane.gov.pl/application
14 https://mojepanstwo.pl/
15 https://mojepanstwo.pl/api
16 Dokumentacja API jest dostępna również w formie dokumentu **[55]**.
17 https://mojepanstwo.pl/kto_tu_rzadzi

- Prawo[18] – serwis prezentuje dostęp do historii publicznie ogłoszonych aktów prawnych i pozwala na wyszukiwanie tekstowe w zbiorach: Dziennika Ustaw, Monitora Polskiego, a także zbiorów prawa lokalnego i kwestii urzędowych.
- Sejmometr[19] – wyszukiwarka wypowiedzi polityków i aktualnych prac Sejmu RP (rys. **3.5**). Ffunkcjonalność pozwala na śledzenie projektów ustaw, wypowiedzi polityków i historii ich głosowań na przestrzeni czasu. Uzupełnieniem powyższego serwisu jest przeglądarka działalności mediów i polityków oraz samorządowców w całej Polsce udostępniana w zakładce 'media' serwisu mojepanstwo[20].

RYSUNEK 3.5. **WIDOK GŁÓWNY Z SERWISU „SEJMOMETR" DLA POSIEDZEŃ SEJMU**



Źródło: https://sejmometr.pl/sejm/posiedzenia

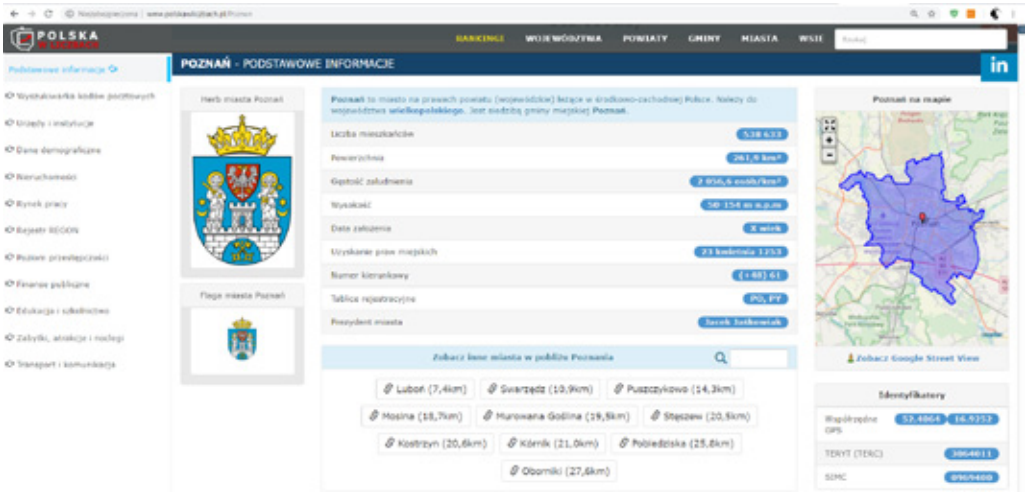
18 https://mojepanstwo.pl/prawo?q=prawo i https://mojeprawo.io/
19 https://sejmometr.pl/
20 https://mojepanstwo.pl/media

Serwis **polskawliczbach**[21] – bardzo dobrze przygotowany serwis wskazujący na kompleksowe opracowania statystyczne dla Polski oraz jednostek samorządowych. Serwis przedstawia wyczerpujące statystyki nawet dla mniejszych gmin i miast w Polsce. Opisuje on m.in.:

- demografię jednostki – strukturę wieku, wykształcenia i dynamikę populacji,
- rynek nieruchomości i rynek pracy, w tym dane dotyczące wynagrodzenia i osób dojeżdżających oraz struktury działalności prowadzonej na jej terenie,
- dane dotyczące przestępczości, finansów, edukacji i wydarzeń kulturalnych realizowanych w jednostce,
- statystyki transportu i zarejestrowanych pojazdów (w tym ich wieku).

Interfejs serwisu został przedstawiony na rys. **3.6**. Strona została zaprojektowana w formie autogenerowanego raportu, który jednak pozwala na efektywne przeglądanie poszczególnych statystyk z wykorzystaniem menu w lewej stronie interfejsu. Przy większości statystyk jasno podane jest źródło i data jego uzyskania. Dodatkowo serwis w automatyczny sposób generuje opis, który tekstem przedstawia najważniejsze informacje o danej jednostce. Przykład dla transportu, wskazujący na szeroki zakres analizowanych danych, zaprezentowany jest na rysunku **3.7**.

RYSUNEK 3.6. **SERWIS „POLSKA W LICZBACH" – PRZYKŁAD OPISU DLA POZNANIA**



Źródło: http://www.polskawliczbach.pl/Poznan

21 http://www.polskawliczbach.pl/

RYSUNEK 3.7. **PRZYKŁAD UTWORZONEGO AUTOMATYCZNIE TEKSTU Z DANYMI STATYSTYCZNYMI W SERWISIE „POLSKA W LICZBACH" – TRANSPORT DLA POZNANIA**

Źródło: http://www.polskawliczbach.pl/Poznan

POZNAŃ - TRANSPORT I KOMUNIKACJA

TRANSPORT I KOMUNIKACJA W PIGUŁCE

Według danych z 2016 roku w wyniku **429** wypadków drogowych w Poznaniu odnotowano **12** ofiar śmiertelnych oraz **521** osób rannych. Oznacza to, że na 100 tys. ludności przypadło **79,4** wypadków *(znacznie więcej od wartości dla województwa wielkopolskiego oraz mniej od wartości dla całej Polski)* oraz **2,2** ofiar śmiertelnych *(znacznie mniej od wartości dla województwa i znacznie mniej od wartości dla kraju)*. Natomiast na 100 tys. pojazdów przypadło odpowiednio **2,7** ofiar śmiertelnych *(znacznie mniej od wartości dla województwa wielkopolskiego oraz znacznie mniej od wartości dla całego kraju)* oraz **115,1** rannych *(znacznie więcej od wartości dla województwa i znacznie mniej od wartości dla Polski)*.

W 2016 roku w Poznaniu zarejestrowanych było **452 628** pojazdów samochodowych oraz ciągników, w tym **356 788** samochodów osobowych (**660,3** na każdy 1000 mieszkańców - *zdecydowanie więcej od wartości dla województwa wielkopolskiego oraz znacznie więcej od wartości dla całej Polski*), **62 535** samochodów ciężarowych (**137,7** - *znacznie więcej od wartości dla województwa oraz znacznie więcej od wartości dla całego kraju*), **1 189** autobusów (**2,2** - *mniej od wartości dla województwa wielkopolskiego oraz znacznie mniej od wartości dla Polski*), **11 856** ciągników siodłowych (**21,9** - *znacznie więcej od wartości dla województwa oraz znacznie więcej od wartości dla kraju*) oraz **14 957** motocykli (**27,7** - *znacznie mniej od wartości dla województwa wielkopolskiego oraz znacznie mniej od wartości dla całej Polski*).

Biorąc pod uwagę najliczniejsze przedziały dla cech zdefiniowanych poniżej:

- typowe auto osobowe ma szacunkowy wiek **14,7 lat**, ma masę całkowitą **1900 kg i więcej**, napędzane jest silnikiem o pojemności **1400-1999 cm3**, stosowane paliwo to **benzyna**.
- typowe auto ciężarowe ma szacunkowy wiek **13,1 lat**, ma ładowność **do 999 kg**, stosowane paliwo to **olej napędowy**.
- typowy autobus ma szacunkowy wiek **14,5 lat**, a jako paliwo stosuje **olej napędowy**.
- typowy ciągnik siodłowy ma szacunkowy wiek **6,4 lat**, a jako paliwo stosuje **olej napędowy**.
- typowy motocykl ma szacunkowy wiek **19,7 lat**.

W 2017 roku w Poznaniu znajdowało się **175** km ścieżek rowerowych, **15,3 km** bus-pasów i **0** parkingów w systemie Parkuj i Jedź (Park & Ride). Zarejestrowano **2 725** taksówek oraz **2 737** licencji na taksówki.

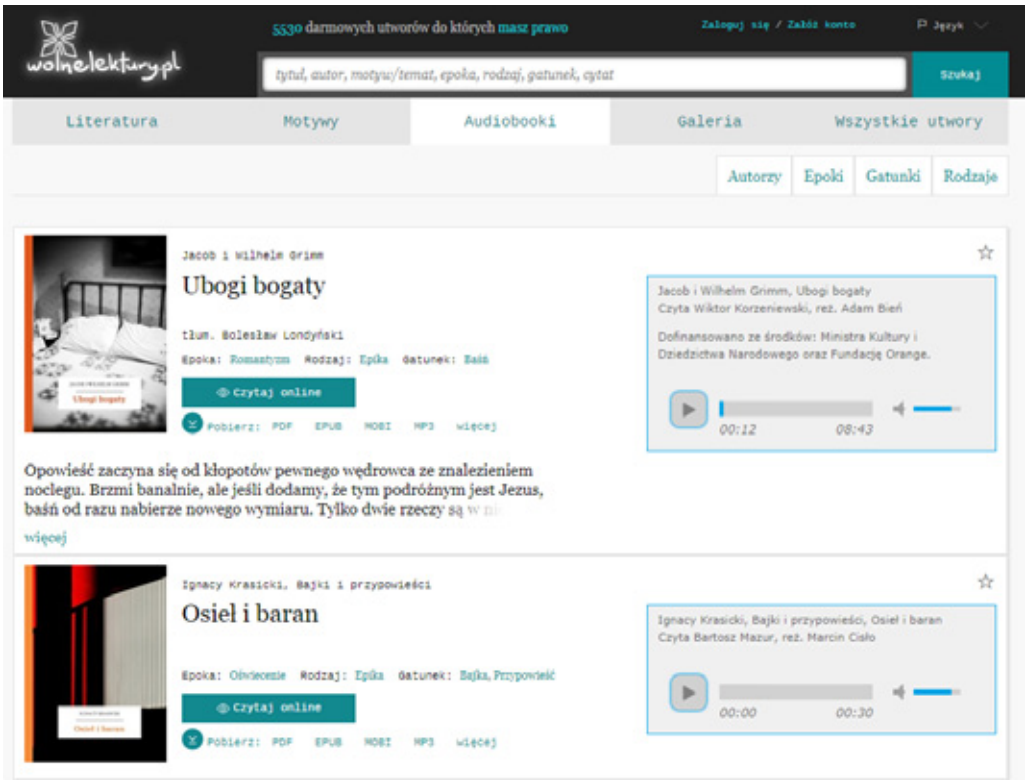
Serwis **wolnelektury**[22] – w tematyce kultury dostępne jest bardzo ciekawe API serwisu wolnelektury. Serwis jest projektem fundacji Nowoczesna Polska i udostępnia informacje o: autorach i dziełach, które są dostępne jako otwarte zbiory danych. Serwis jest bardzo rozbudowany i pozwala m.in. na udostępnianie fragmentów konkretnych utworów w sposób zautomatyzowany (przykład https://wolnelektury.pl/api/books/romeo-ijulia/fragments/1189496159453/. Serwis udostępnia nie tylko książki i audiobooki, ale także pozwala na przeglądanie motywów literackich, rodzajów literatury, epok czy autorów. Przykład strony jednej z lektur zawarty jest w listingu **3.6**.

22 https://wolnelektury.pl/api/

KOD ŹRÓDŁOWY 3.6. **PRZYKŁAD WYNIKU API SERWISU WOLNELEKTURY DLA JEDNEJ Z KSIĄŻEK**

```
{
    "kind": "Dramat",
    "full_sort_key": "mickiewicz adam$dziady$830",
    "title": "Dziady",
    "url": "https://wolnelektury.pl/katalog/lektura/dziady/",
    "has_audio": false,
    "author": "Adam Mickiewicz",
    "cover": "book/cover/dziady.jpg",
    "slug": "dziady",
    "epoch": "Romantyzm",
    "href": "https://wolnelektury.pl/api/books/dziady/",
    "genre": "Dramat romantyczny",
    "simple_thumb": "https://wolnelektury.pl/media/book/cover_api_thumb/dziady.jpg",
    "cover_color": "\#db4b16",
    "cover_thumb": "book/cover_thumb/dziady_6c9vXoz.jpg"
}
```

RYSUNEK 3.8. **SERWIS WOLNE LEKTURY – KATALOG AUDIOBOOKÓW**



Źródło: https://wolnelektury.pl/katalog/audiobooki/

widok.gov.pl

W przypadku dynamicznie rozwijającego się katalogu zasobów danych (w formie usług, stron internetowych i zbiorów danych), warto również wskazać bardzo ciekawą nową inicjatywę powstałą jako realizacja fragmentu funkcjonalności Zintegrowanej Platformy Analitycznej[23], za którą odpowiedzialne jest Ministerstwo Cyfryzacji.

Portal udostępnia przede wszystkim statystyki wykorzystania poszczególnych usług administracji publicznej (podobnie do brytyjskiego przykładu)[24]. Z najciekawszych elementów pokazuje on:

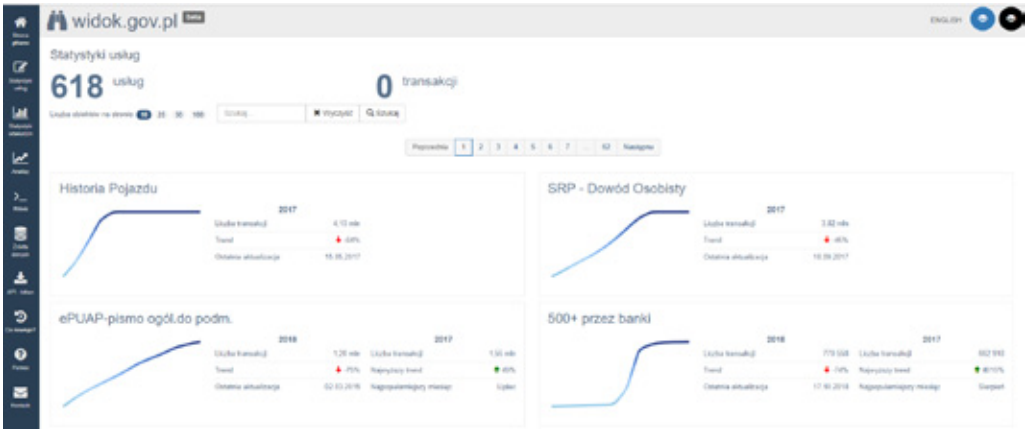
- statystyki odwiedzin polskich portali i usług administracyjnych względem: fraz wyszukiwania, platform i urządzeń dostępu, przeglądarek i systemów operacyjnych,
- statystyki wykorzystania usług publicznych, które służyć mogą również za ich katalog,
- analizy przeprowadzone przez twórców portalu.

Portal jest ciekawą inicjatywą z perspektywy raportu, głównie ze względu na to, że tworzy **swoisty katalog usług publicznych**, a także informuje o ich **popularności**. Analizy na nim umieszczone mogą służyć dwóm podstawowym celom:

- dla obywateli – pozwala na identyfikację samych usług i wskazuje podatnikom "za co płacą" w perspektywie cyfryzacji usług administracyjnych,
- dla administracji – wskazują na popularność poszczególnych usług i informują o tym, jakie dane powinny być brane pod uwagę w procesie analizy wykorzystania i opracowywania nowych usług rządowych.

23 https://www.gov.pl/web/cyfryzacja/zintegrowanaplatforma-analityczna
24 https://www.gov.uk/ performance/services

RYSUNEK 3.9. **SERWIS WIDOK.GOV.PL PRZEDSTAWIAJĄCY STATYSTYKI WYKORZYSTANYCH POLSKICH USŁUG PUBLICZNYCH**



Źródło: https://widok.gov.pl/services/

Ten portal internetowy to ciekawa inicjatywa, mająca na celu zagregowanie dostępnych w polskiej administracji usług, wraz z platformą analityczną. Co ciekawe, uruchomiona funkcja RWeb[25] wydaje się bardzo ciekawą perspektywą dostępu do danych publicznych (np. o liczbie odwiedzin). Potraktowanie tej usługi jako inspiracji, daje podstawy do myślenia w przyszłości o możliwości rozszerzenia usług udostępniających dane publiczne także o środowisko pozwalające chociażby na tworzenie raportów z dostępnych danych źródłowych. Otwiera to nowe możliwości interakcji z platformami udostępniającymi dane, w celu zapisu i dzielenia się analizami w oparciu o taki interfejs analityczny.

Portal, niestety, mimo świetnego wykonania technicznego i wizualnego, sam w sobie jest mało popularny i powinien być częścią jakiejś większej struktury. Podsumowując, inicjatywa jest słuszna z perspektywy otwartości danych i przydałaby się jej integracja z portalem dane.gov.pl, np. jako nowa zakładka, wraz z poszerzeniem tematyki identyfikacji usług dynamicznych (API). Poszerzenie portalu o środowisko analityczne i umieszczenie przykładowych dynamicznie generowanych analiz pozwoliłoby także na zapoznanie administracji rządowej i samorządowej z możliwościami wykorzystania danych wewnętrznych do poprawy świadczonych przez nich usług.

25 https://widok.gov.pl/rweb/

Usługi komercyjne

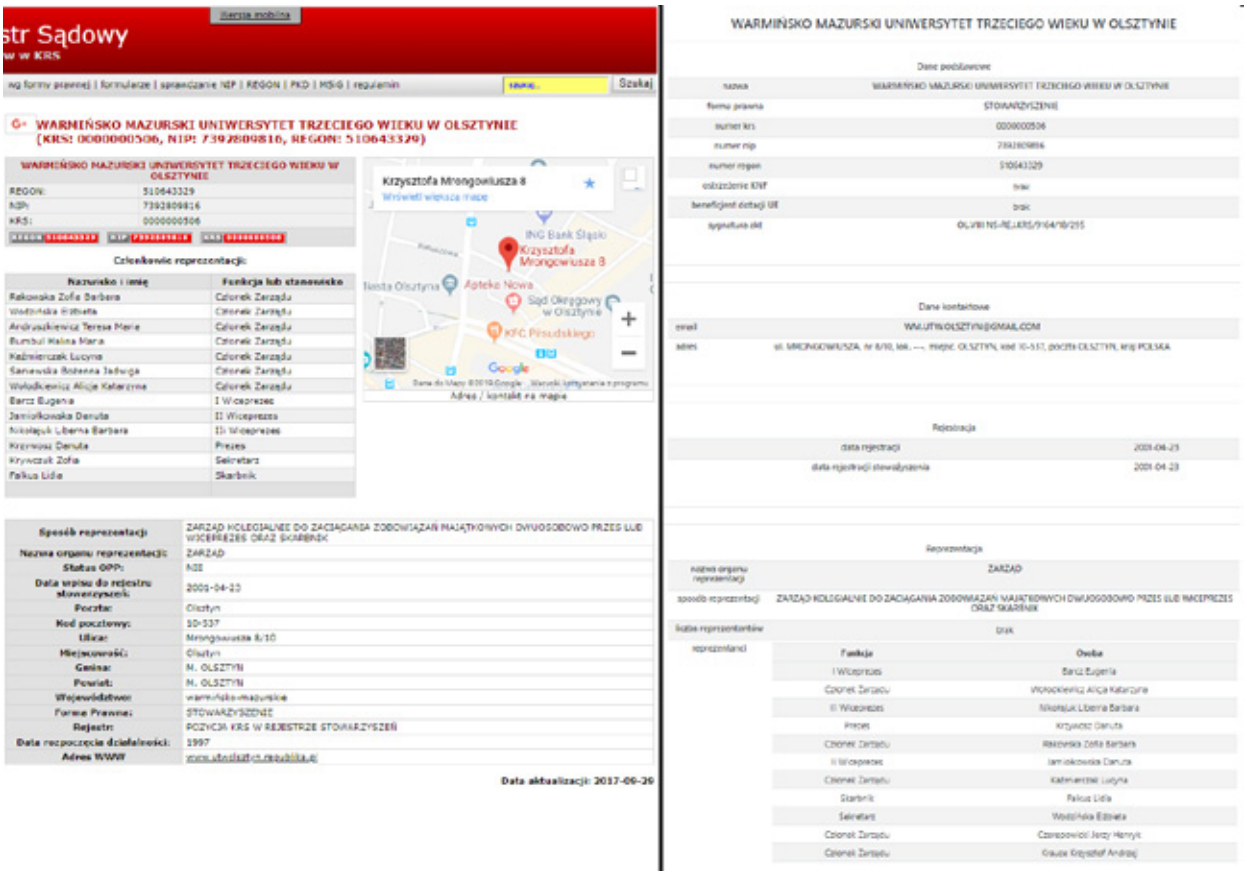
Wśród przykładów przedsiębiorców, którzy zbudowali swoje komercyjne usługi na podstawie otwartych zasobów danych (i bazują swój model biznesowy na przetwarzaniu, automatyzacji i udostępnianiu danych) wskazać można kilka przykładów.

- Dla zbiorów REGON, KRS, CEIDG i innych informacji o przedsiębiorcach rejestr.io, http://www.krs-online.com.pl/, https://filtrkrs.pl, http://www.informacjakredytowa.pl, https://transparentdata.pl/rozwiazania/api/, https://przeswietl.pl/. Serwisy te spełniają podobną rolę, ich porównanie dla losowo wybranego rekordu jest zaprezentowane na rysunku **3.10** i **3.11**. Analizując treść przedstawionego tam przykładu można zauważyć, że zawarte są różne informacje kontaktowe: w jednym przypadku strona internetowa oraz informacje o aktualności podanego wpisu. Dodatkowo usługi rejestr.io i przeswietl.pl w opcji premium udostępniają informacje o wierzytelnościach, zaległościach, postępowaniach upadłościowych i długach, a także obecności podmiotu m.in. w rejestrze dłużników niewypłacalnych czy publicznych giełdach długów. Usługa prześwietl.pl udostępnia API pozwalające na przetworzenie wszystkich tych informacji w sposób automatyczny, celem np. zasilenia firmowego systemu. Serwis informacjakredytowa posiada API w standardzie formatu wymiany danych XML, jednak nie wydawała się być otwarcie dostępna online.
- **Open Data Whiteaster**[26] i https://www.otwartedane.com/. Spółka Whiteaster specjalizuje się m.in w zapewnianiu platform open data w spójny sposób dla jednostek samorządowych. Posiada w tej chwili listę zbiorów publicznych i udostępnianych na podstawie umów z poszczególnymi dostawcami. Firma pozwala na dostosowanie udostępniania powierzonych zbiorów poprzez interfejs API (przykład na rys. **3.12**).
- **Lex.pl**[27] (Wolters Kluwer). Serwis ten to wyspecjalizowany system pozwalający na analizę i przeszukiwanie aktów prawnych. Zbudowany jest w większości w oparciu o dostępne zbiory i autorski system analizy tekstu. Pozwala m.in. na wyszukiwanie konkretnych fraz w orzeczeniach i dokumentach prawnych.
- **Biuro Informacji Gospodarczej InfoMonitor**[28] będący przedsiębiorstwem i jednocześnie inicjatywą wielu przedsiębiorstw i spółek państwowych w celu prowadzenia ujednoliconej bazy przechowującej i udostępniającej informacje gospodarcze o zadłużeniu oraz terminowych płatnościach osób i przedsiębiorców. Skorzystanie z serwisu wymaga rejestracji.

26 https://opendata.whiteaster.com/dataset
27 http://www.lex.pl
28 https://www.big.pl/

- **Airly**[29] – to przedsiębiorstwo założone początkowo przez studentów AGH, które opracowało własny czujnik mierzący pyły zawieszone PM10 i PM2.5, temperaturę, ciśnienie i wilgotność. W tej chwili za pomocą sieci czujników udostępniają dane o jakości powietrza w podobny sposób jak serwis GIOS (API IMGW), posiadają ponad 1000 czujników zamontowanych w całej Polsce. Usługa udostępnia dane za pomocą API oraz interaktywnej mapy (pokazanej na rys. **3.13**). Istnieją też podobne usługi, np. **Looko2**[30].

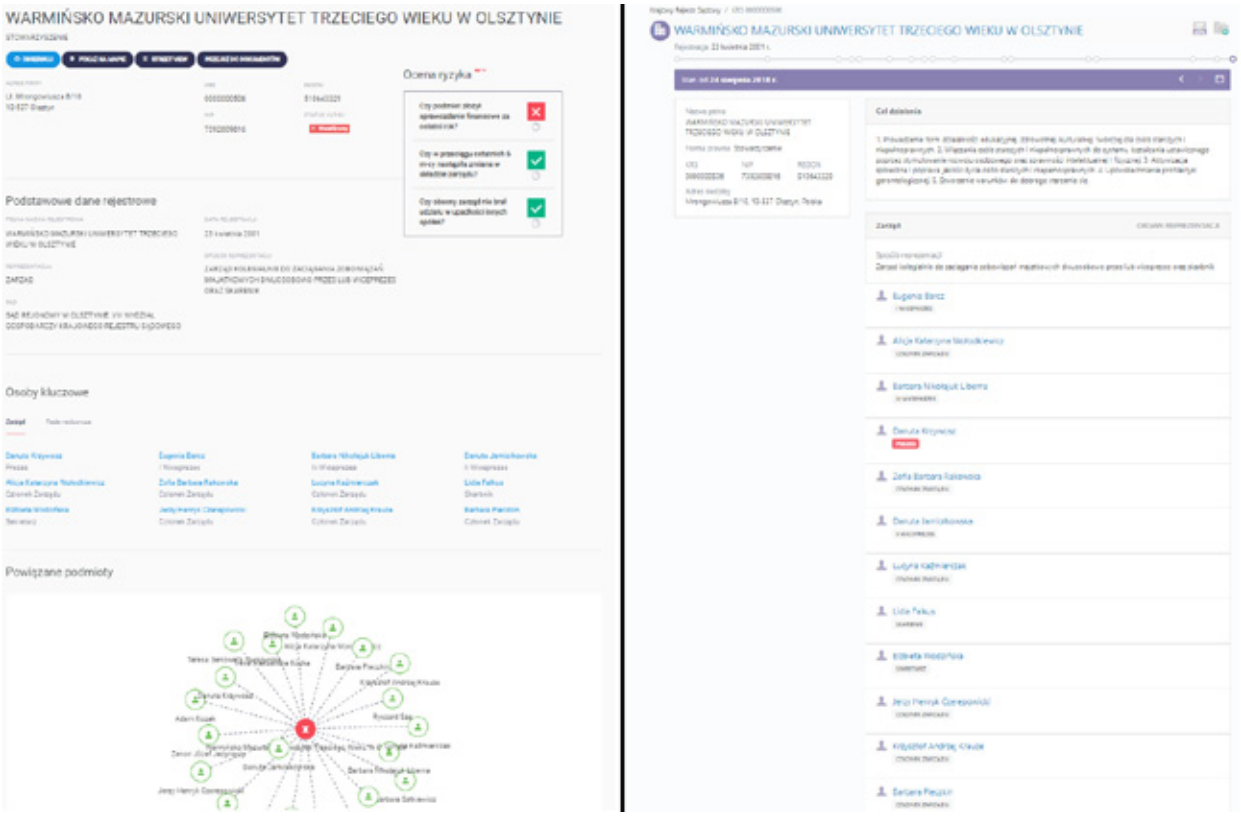
RYSUNEK 3.10. **PORÓWNANIE SERWISÓW ZAWIERAJĄCYCH INFORMACJE O PRZEDSIĘBIORCACH**



Źródło: opracowanie własne na podstawie http://www.krs-online.com.pl/ i https:// filtrkrs.pl/.

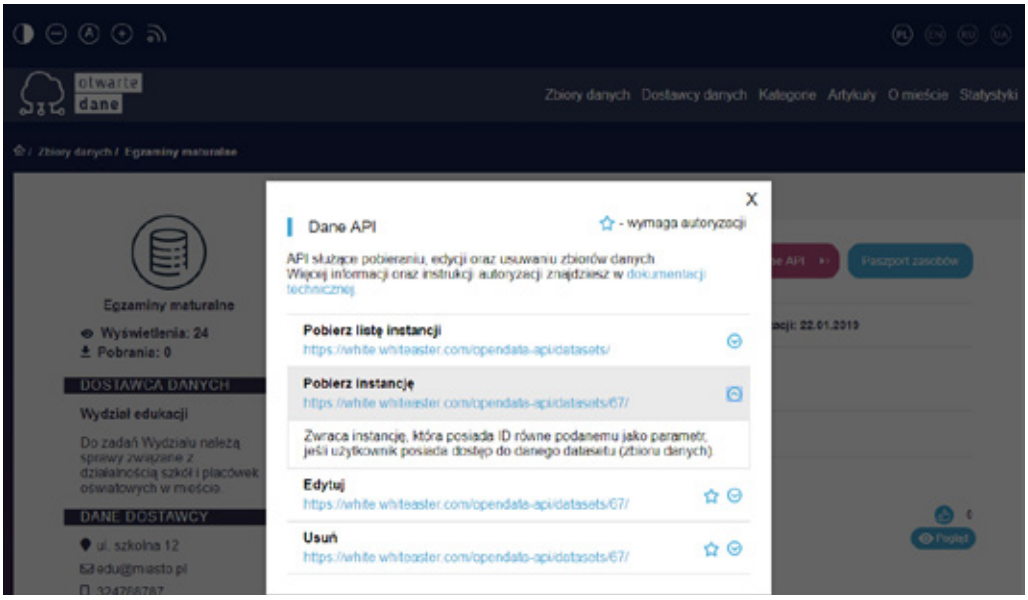
29 https://airly.eu/map/pl/
30 http://www.looko2.com/

RYSUNEK 3.11. **PORÓWNANIE SERWISÓW ZAWIERAJĄCYCH INFORMACJE O PRZEDSIĘBIORCACH**



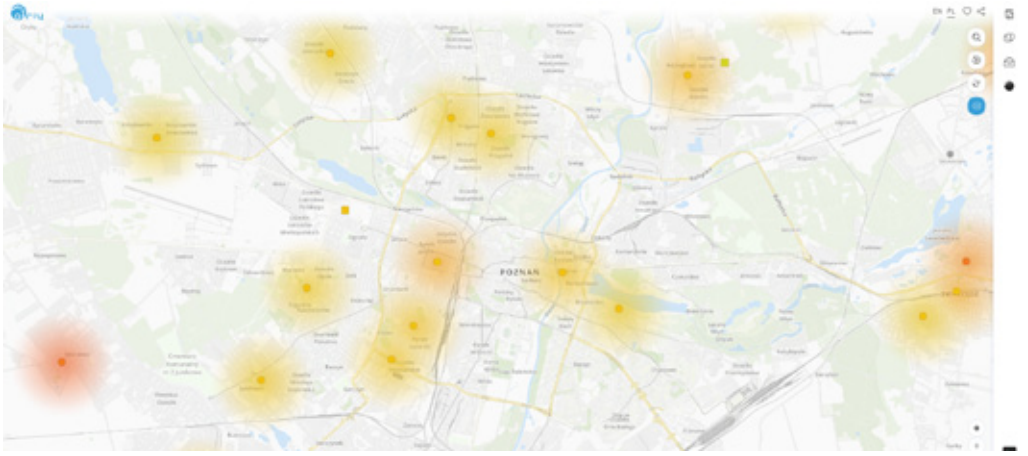
Źródło: opracowanie własne na podstawie: https://przeswietl.pl/ i rejestr.io.

RYSUNEK 3.12. **PRZYKŁAD ZBIORU UDOSTĘPNIONEGO NA PLATFORMIE FIRMY WHITEASTER WRAZ Z DOSTĘPEM DO API**



Źródło: https://opendata.whiteaster.com/dataset

RYSUNEK 3.13. **MAPA AIRLY POKAZUJĄCA STAN ZANIECZYSZCZENIA POWIETRZA DLA POZNANIA W SERWISIE ONLINE**



Źródło: https://airly.eu/map/pl/

3.2.3. Samorządy lokalne

W przypadku samorządów lokalnych także realizowane są działania w obszarze udostępniania danych zgodnie z ideą open data. Charakterystyka danych udostępnionych przez wybrane samorządy jest zaprezentowana w tabeli **3.1**. Przede wszystkim miasta udostępniają wiele podstawowych zbiorów danych o statystykach związanych geograficznie z obszarem miasta. **Problemem jest często aktualność udostępnionych danych**. Przeanalizowanie wszystkich zbiorów i usług udostępnionych na wszystkich platformach open data wykraczałoby poza zakres niniejszego raportu. Dlatego celem tego podrozdziału jest opisanie ogólnych dobrych praktyk i kilku często zaobserwowanych błędów, na które trzeba zwrócić uwagę przy korzystaniu z tych zbiorów.

Jeden z najlepszych portali otwartych danych prowadzony jest przez **Gdańsk**, udostępniając w większości aktualne dane, w wielu formatach, w tym duży zbiór danych geograficznych. Na pochwałę zasługuje przede wszystkim API do systemu kolejkowego w odpowiednich działach urzędu pozwalający na automatyczne przetwarzanie tych danych. Usług nie jest wcale zbyt wiele, ale są dobrze przygotowane. Wrażenie może popsuć brak nazwy domeny w serwisie firmy tristar, które udostępnia dane ZTM Gdańsk, co może budzić wątpliwości co do stanu wersji produkcyjnej usługi. Gdańsk został wyróżniony przez udostępnione przez siebie usługi i zbiory w raporcie Europejskiego Portalu Danych w roku 2017 [**20**].

Wrocław posiada bardzo dobrą platformę do udostępniania danych, która charakteryzuje się największym poziomem zaawansowania technicznego, jednak dane w niej umieszczone wydają się być raczej testowo, często z datami wskazującymi na manualną i rzadką aktualizację. Warto pochwalić uwzględnienie standardów open data, jednak aktualność umieszczonych danych można poddać w wątpliwość. Co ciekawe, dane publikowane przez Zarząd Dróg i Utrzymania Miasta we Wrocławiu charakteryzują się raczej dobrym systemem aktualizacji.

Poznań udostępnia za to małą liczbę usług, ale są to wszystko dobrze przemyślane usługi udostępnione online. W szczególności warta uwagi jest najbardziej rozbudowana platforma dostępu do geodanych (w porównaniu do innych miast) i wykorzystanie w niej formatu GeoJson. Dostępne jest też wiele przydatnych usług, które są udostępniane w formie map czy interaktywnych prezentacji, co bardzo utrudnia wykorzystanie i budowanie usług na ich podstawie (nie jest to jednak niemożliwe).

Warszawa udostępnia bardzo dobre usługi związane z pozycją tramwajów i autobusów (zdaniem autorów raportu jest to najlepsza usługa tego typu), jednak nawigacja po stronie jest bardzo skomplikowana: strona zawiera wiele interaktywnych dokumentów, brak referencyjnych URL dla poszczególnych zbiorów praktycznie uniemożliwia wygodne korzystanie z dobrze skompletowanych zbiorów danych. Miejskie centrum kontaktu 19115 to przykład dobrej usługi,

która, co najważniejsze dla automatycznego przetwarzania, pozwala także na przeglądanie zgłoszeń za pomocą API.

Gdynia natomiast udostępnia najwięcej zbiorów danych, co może stanowić podstawę do weryfikacji, czy tego rodzaju agregaty są też zbierane przez inne podmioty samorządowe (i mogą zostać wykorzystane). Jednak przez ogrom zbiorów wiele z nich wydaje się być aktualizowanych z opóźnieniem. Ciekawe i aktualizowane są zbiory utrzymywane przez Zarząd Dróg i Zieleni, choć droga do zasobu mogłaby być krótsza, a opisy dokładniejsze.

TABELA 3.1. **CHARAKTERYSTYKA OTWARTYCH ZBIORÓW DANYCH DLA WYBRANYCH MIAST W POLSCE**

MIASTO	API	TEMATY	PRZYKŁAD CIEKAWYCH USŁUG DOSTĘPNYCH AUTOMATYCZNIE	AKTUALNOŚĆ DANYCH
Gdańsk[31] (60 zbiorów)	Ograniczone (niektóre usługi)	Edukacja (5) Bezpieczeństwo (1) Kultura (4) Dane przestrzenne (2) Demografia (8) Gospodarka (10) Urząd Miejski (11) Zdrowie (2) Środowisko (7) Sport i rekreacja (5) Transport (5)	**System GIS** https://www.gis.gdansk.pl/indexipg.html. **Kalendarz wydarzeń** https://planerkulturalny.pl z dostępnym REST API. **Poziom wody i opadów** https://pomiary.gdanskiewody.pl/rest. **Aktualny stan odjazdów na przystankach ZTM Gdańsk** https: //91.244.248.30/dataset/tristar (pozwalający na dynamiczne sprawdzanie opóźnień na przystankach https://87.98.237.99: 88/delays?stopId=34140). **System kolejek w urzędach** (API XML) https://www.gdansk.pl/otwartedane?show=data&id=systemykolejkowe-w-zespolachobslugi-mieszkancow. **Stan wykorzystania ścieżek rowerowych dla Gdańska** (wymaga scrappingu) http://www.rowerowygdansk.pl/start,169,170.html.	Tak, dla usług udostępnianych API.
Wrocław[32] (47 zbiorów)	Tak, do wszystkich zasobów	Edukacja (2) Dane przestrzenne (10) Demografia (1) Inne (2) Sport i rekreacja (2) Wydarzenia (2) Urząd Miejski (10) Środowisko (2) Sprawy społeczne (1) Transport (14)	Dane dodatkowo spełniają wymogi dla 4 poziomu otwartych danych (formaty Linked Data). **Mapa pozycji pojazdów** (wymagany scrapping) https://mpk.wroc.pl/ jak-jezdzimy/mapa-pozycjipojazdow. **Stan parkingów, pozycji komunikacji miejskiej i natężenia ruchu** https://www.wroclaw.pl/opendata/organization/zarzaddrog-i-utrzymania-miasta-wewroclawiu.	Nie, dane wydają się nie być aktualizowane i być umieszczone na serwisie testowo.

31 https://www:gdansk:pl/otwarte-dane
32 https://www.wroclaw.pl/open-data/

MIASTO	API	TEMATY	PRZYKŁAD CIEKAWYCH USŁUG DOSTĘPNYCH AUTOMATYCZNIE	AKTUALNOŚĆ DANYCH
Poznań[33]	Tak	Dane przestrzenne Transport Wydarzenia[34]	**Informacje na temat wydarzeń i aktualności** https://www.um.poznan.pl/ web-service/events?wsdl. **Stan Parkingów** https: //www.ztm.poznan.pl/pl/dladeweloperow/ parkingFiles. **Stan kolejek w urzędzie** (wymaga Scrappingu) https://www.um.poznan.pl/ mim/ rezerwacje/?co=kolejki. **API dla danych przestrzennych i obiektów zabytkowych** https://egov.psnc.pl/node/29#rest_ api_dla_danych_ przestrzennych. **Aktualny stan odjazdów na przystankach ZTM Poznań** (usługa skomplikowana do Scrappowania) https://www.peka.poznan.pl/vm/.	Tak (wszystkie zbiory udostępnione jako usługi).
Warszawa[35] (ok. 50 zbiorów)	Tak- wymaga rejestracji i długiej nawigacji po serwisie.	Historia i Religia Bezpieczeństwo i interwencje Transport miejski Kultura Edukacja Ekologia Sport i rekreacja Dane urzędowe Dane online Nieruchomości Projekty społeczne Dane przestrzenne Inne	**Ciekawe usługi udostępnione w kategorii Dane online.** **Aktualna pozycja autobusów i tramwajów.** **Stan kolejek w urzędach.** **Miejskie Centrum Kontaktu - 19115.** **Zgłaszanie i przeglądanie usterek.**	Tak.
Gdynia[36] (ponad 170 zbiorów)	Tak- prawie wszystkie zbiory są dostępne w API	Transport (47) Bezpieczeństwo (30) Edukacja (18) Woda i usługi sanitarne (10) Odpady stałe (8) Wartości odniesienia (8) Zdrowie (7) Administracja (6) Budownictwo (6) Gospodarka (5) Środowisko (5) Bezrobocie (4) Finanse (3) Gospodarka morska (3) Ludność (3) Podmioty gospodarcze (3) Telekomunikacja i innowacje (3) Energia (2) Planowanie przestrzenne (2) Schroniska (2) Kultura (1)	**Dane z systemu transportu dostępne są po pobraniu specyfikacji dla systemu TRISTAR** https://otwartedane.gdynia.pl/ pl/dataset/ dokumentacjatristar-api. **Pomiary prędkości na segmentach dróg** http://api.zdiz.gdynia.pl/ ri/rest/traffic_speed_ data?segmentId=33737.	Nie- niektóre dane są z roku 2015, niektóre są publikowane z miesięcznym opóźnieniem. Większość danych jest aktualizowana ręcznie z dużym opóźnieniem.

Źródło: opracowanie własne.

33 https://www.poznan.pl/mim/api/
34 Poznań udostępnia jedynie dynamiczne usługi jako API, dlatego brak liczb przy liczbie zbiorów.
35 https://api.um.warszawa.pl/
36 https://otwartedane.gdynia.pl

Pewnym problemem w publikacji danych wydaje się nadmierne dzielenie zbiorów danych i poziom agregacji zmiennych. Być może z powodu obaw przed niezachowaniem reguł anonimizacji na wielu serwisach można spotkać m.in „kwartalne statystyki dot. przestępczości" w poszczególnych kategoriach. Wydaje się, że zgodnie z ogólnymi założeniami można by było udostępnić mniej zagregowany zbiór (np. tygodniowe, dzienne interwencje) wraz z kategoryzacją poszczególnych przestępstw, co stanowczo poszerzyłoby zastosowania udostępnionych danych.

Przykłady międzynarodowe

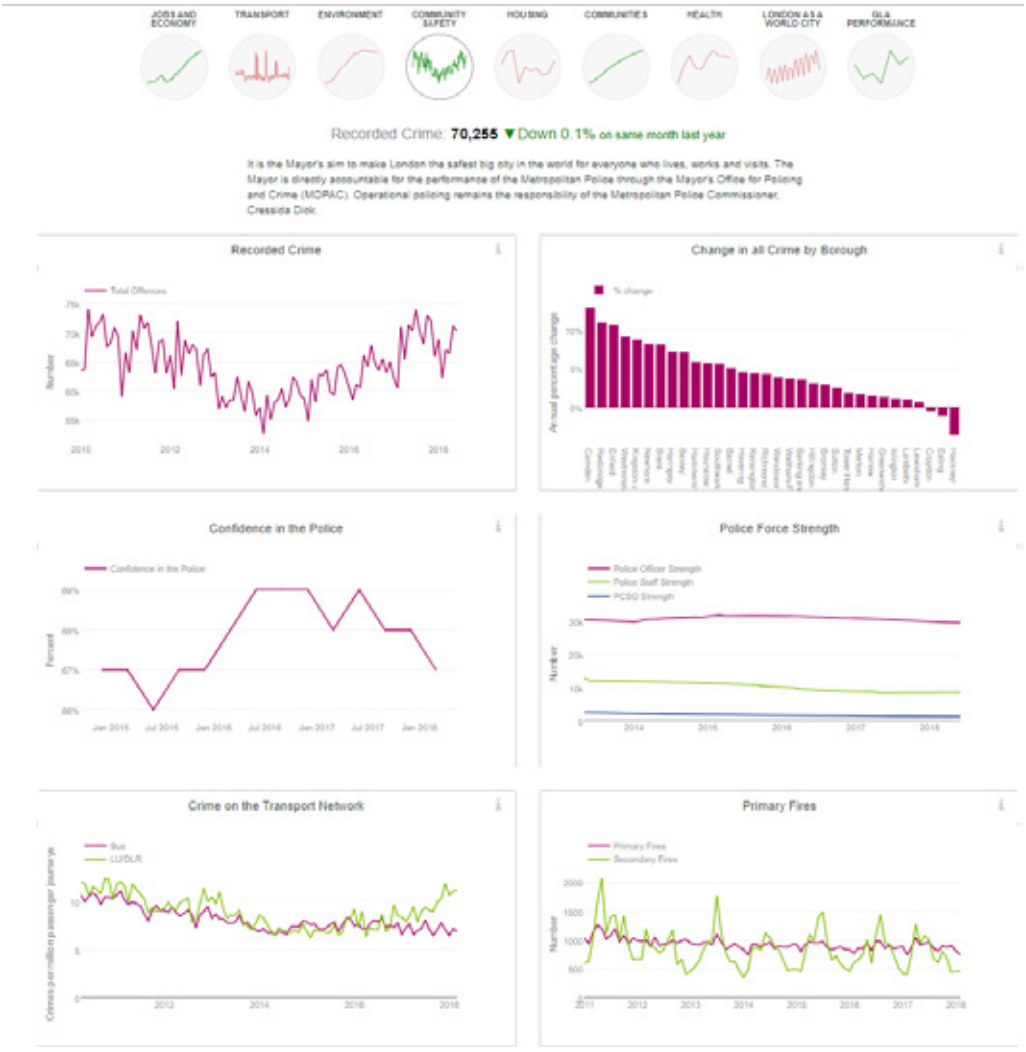
Na podstawie powyższych przykładów udostępniania zbiorów danych można przyrównać je do światowych liderów prowadzących podobne portale. Przykładowo portal dla Londynu[37] pokazuje najważniejsze miejskie statystyki (dot. transportu, ochrony środowiska, rynku pracy, czy raportów zdrowotnych) w stosunku do okresów poprzednich (rys. **3.14**). Pozwala to na ocenę stanu i rozwoju miasta bezpośrednio dla inwestorów i daje wgląd w działania administracji dla obywateli miasta. Umożliwia też interaktywny dostęp do wszystkich danych statystycznych przypisanych do konkretnego rejonu, a także na wylistowanie wszystkich zbiorów danych za pomocą krótkiego linku udostępnionego w API[38]. Barcelona[39] posiada serwis o strukturze bardzo podobnej do wyżej wymienionych polskich serwisów, do każdego zbioru danych posiada jednak informację o tym, czy zbiór danych jest aktualny i przypisaną kategorię częstotliwości aktualizacji. Aż 42 zbiory aktualizowane są na bieżąco i pokrywają one m.in. stan parkingów, stacji rowerowych, środowiska i jakości powietrza oraz usługi wykorzystujące mapy i system WMS.

Dane o stacjach rowerowych, wypadkach (z udziałem pojazdów, pieszych, rowerów) udostępnia wiele miast m.in Barcelona czy Nowy Jork[40]. To drugie z miast prezentuje m.in w spójny sposób (z dostępem przez API) aktualne oferty pracy, przetargi i publiczne licytacje, aktualne projekty budowlane i utrudnienia na drogach. Melbourne[41] natomiast wykorzystuje wiele danych zbieranych z autonomicznych sensorów, dając pełny obraz stanu transportu i „życia" miasta, wykorzystując czujniki: stacji rowerów i parkingów, natężenia ruchu samochodów, rowerów i pieszych, pogody i stanu powietrza, wraz z danymi geograficznymi uwzględniającymi stan zieleni i kategorię zabudowy w poszczególnych rejonach miasta. Przykładów innych miast posiadających szerokie serwisy open data można znaleźć wiele (m.in. w artykule Forbes [**116**]). Zapoznanie się z nimi może wskazać najlepsze przykłady realnie wykorzystywanych przez obywateli i przedsiębiorców usług, co może w pozytywny sposób wspomóc w przyszłości rozwój

37 https://data.london.gov.uk
38 https://data.london.gov.uk/api/action/package_list
39 http://opendata-ajuntament.barcelona.cat/data/en/dataset
40 https://opendata.cityofnewyork.us/
41 https://data.melbourne.vic.gov.au/

tego sektora. Podsumowując, zbiory danych, usługi i platformy wykorzystywane w celu udostępniania danych publicznych przez samorządy lokalne wśród czołowych liderów nie różnią się znacząco od zbiorów danych udostępnionych przez polskich usługodawców. Na podstawie analizy wskazać można standard usług, których warto byłoby się spodziewać niedługo w większych polskich miastach w kategoriach obejmujących:

RYSUNEK 3.14. **PRZYKŁAD WIZUALIZACJI UŻYTECZNYCH STATYSTYK W SERWISIE OPEN DATA DLA LONDYNU**



Źródło: https://data.london.gov.uk/

- Transport – lokalizację pojazdów transportu publicznego (na podstawie GPS), a także stacji rowerowych i danych z sensorów.
- Statystykę – szczegółowych niezagregowanych danych (np. na poziomie dzielnicy) dotyczących, w jak najmniejszych agregatach czasowych: przestępczości w podziale na kategorie, rynku pracy.
- Rejestrów – dotyczących prac konserwacyjnych, przetargów i spraw (ogłoszeń) publicznych, a także wydarzeń organizowanych na terenie objętym działaniem systemu.
- Spraw urzędowych – uwzględniających system kolejek do większości działów urzędów i instytucji publicznych (być może z systemem automatycznego zapisu przez API – co pozwoliłoby na zbudowanie aplikacji mobilnej), a także wszelakiego rejestru pozwoleń i wydawanych licencji, kontroli i podejmowanych działań (np. przebudowy publicznego budynku lub projektów inwestycyjnych).
- System map i WMS – jako spójny system uwzględniający wszystkie warstwy, w tym: charakterystykę zabudowy, aktualny stan i zgłaszane problemy oraz w dynamiczny sposób obejmujący decyzje administracyjne. System wraz z dobrym wzorcem opisu, który na to pozwala, został już zbudowany i jest zaprezentowany na rysunku **3.19**.

Lista ta jest spójna ze strategiami przyjętymi w zaktualizowanej dyrektywie PSI (podrozdział **3.3.2**). Ciekawe przykłady najnowszych usług w kategorii smart-cities pozwolą w przyszłości do powyższych dołączyć także dane dotyczące zużycia wody, prądu, poziomu hałasu czy mapę tematów poruszanych w mediach społecznościowych. Spodziewać się można dynamicznego rozwoju wyżej wymienionych obszarów i pojawiania się w możliwie jak największej liczbie miast podobnych usług. Jednocześnie przy ich obecnym stanie możliwe jest budowanie usług wykorzystujących już udostępnione zbiory.

3.3. Inicjatywy open data

3.3.1. Inicjatywy światowe

Największą globalną inicjatywą open data jest ta zarządzana przez Organizację Narodów Zjednoczonych (ONZ). W raporcie "World That Counts" grupa doradcza ds. rewolucji w danych dla zrównoważonego rozwoju[42] rekomendowała ustanowienie Światowego Forum Danych ONZ (UN World Data Forum) [**65**]. Pierwsze takie forum miało miejsce w styczniu 2017 roku w Republice Południowej Afryki, a kolejne w Zjednoczonych Emiratach Arabskich w październiku 2018 roku[43]. Światowe Forum Danych ma pełnić rolę katalizatora dla budowania współpracy pomiędzy różnymi grupami ekspertów oraz osób pracujących zawodowo z danymi, m.in. naukowców od danych (data scientists), współtworzących polityki, działaczy społecznych, jak i użytkowników końcowych.

42 Independent Expert and Advisory Group on Data Revolution for Sustainable Development
43 https://undataforum.org/

3.3.2. Europejska dyrektywa PSI

Biorąc pod uwagę aktualną strategię otwierania danych dla Europy nastąpiło uaktualnienie dyrektywy PSI w sprawie ponownego wykorzystywania informacji sektora publicznego. W źródłowym dokumencie [**92**] wskazano m.in. obszary, które zauważono jako kluczowe dla dalszego rozwoju otwartych danych, wskazując:

- Dane generowane poprzez sektor transportowy i usługi transportowe realizowane na terenie miejskim za zgodą samorządów (takie jak parkingi, stacje rowerowe).
- Dynamiczne zbiory danych dostępne przez API – pozwalające na dostęp w czasie rzeczywistym.
- Kwestie otwierania możliwie jak największej liczby zbiorów, w stosunku do których pojawiają się zapytania i wiążą się one w tym momencie z kosztami ponoszonymi przez przedsiębiorcę. Udostępnienie takich danych nieodpłatnie w sposób automatyczny w długim okresie staje się korzystne dla jednostki lokalnej.
- Unikanie umów na wyłączność i pracy jedynie z jednym dostawcą usług w celu udostępnienia danych. Promowanie platform udostępniających dane (np. API z pozycją autobusów) zamiast umów czy przetargów o stworzenie aplikacji udostępniającej określone funkcjonalności.

W związku z powyższymi obszarami zaproponowano rekomendacje, np. nie pobieranie dodatkowych opłat za ponowne udostępnianie już zebranych danych, udostępnienie danych ze wszystkich usług realizowanych za pieniądze publiczne i jednostek świadczących usługi publiczne (w tym badań naukowych), minimalizowanie sytuacji, w której duzi dostawcy i firmy (np. negocjujące warunki udostępnienia danych czy API) mają przewagę na rynku nad niezależnymi twórcami oraz zwiększanie udziału API w tworzonych usługach. Te właśnie strategie powinny przyświecać twórcom usług open data w najbliższym czasie. 22 stycznia 2019 roku zostało podpisane porozumienie, które przyjęło zmienioną dyrektywę uwzględniającą powyższe cele strategiczne. Dyrektywa ta z pewnością wpłynie na sposób tworzenia nowych usług opartych na otwartych zbiorach danych zarówno dla podmiotów ogólnokrajowych, jak i samorządowych.

3.3.3. European Data Portal

Europejski Portal Danych[44] przygotowuje coroczne raporty oceniające dojrzałość poszczególnych państw UE we wdrażaniu open data. Wyniki rankingu w roku 2018 są widoczne na rys. **3.15**.

44 https://www.europeandataportal.eu/pl

RYSUNEK 3.15. **15 PAŃSTW Z NAJWIĘKSZYM WSPÓŁCZYNNIKIEM OPEN DATA MATURITY W 2018 ROKU**

	Policy			Portals				Impact								Quality				
Country	Policy framework	National coordination	Licencing norms	Portal features	Portal usage	Data provision	Portal sustainability	Strategic awareness	Political impact	Social impact	Environmental impact	Economic impact	Automation	Data & metadata currency	DCAT-AP compliance	Policy	Portal	Impact	Data Quality	Open Data Maturity
Maximum	180	350	150	250	120	160	120	200	130	110	80	130	100	210	210	680	650	650	520	2500
Ireland See details	170	310	140	200	100	140	85	175	130	110	80	130	80	150	195	620	625	625	435	87.8%
Spain See details	180	310	145	185	110	125	85	180	130	110	80	130	90	130	135	635	505	630	405	87.0%
France See details	180	320	145	220	115	100	85	160	115	110	80	50	80	135	130	645	529	515	395	83.0%
Italy See details	180	330	145	160	95	115	95	125	120	110	80	40	100	165	150	655	455	475	415	80.0%
Cyprus See details	140	320	150	200	120	145	80	165	115	95	65	15	65	115	200	610	645	455	380	79.6%
Luxembourg See details	140	335	145	185	115	125	70	195	120	90	50	110	55	70	175	620	495	485	380	76.0%
Slovenia See details	180	325	145	170	115	135	85	115	100	80	50	0	70	105	190	650	505	345	360	74.0%
Greece See details	150	275	130	185	110	130	85	170	100	70	40	80	45	120	150	555	513	408	315	73.6%
Slovakia See details	180	295	145	190	95	135	45	165	70	60	50	40	55	125	190	620	485	385	370	73.6%
Netherlands See details	140	315	140	135	120	115	85	115	85	60	50	30	90	155	190	595	455	340	435	73.0%
United Kingdom See details	150	285	115	160	95	90	70	135	100	70	50	50	60	120	170	560	415	445	360	70.8%
Latvia See details	120	270	140	145	85	140	30	165	70	30	50	40	50	140	130	530	405	355	370	66.2%
Poland See details	150	270	130	160	105	135	65	155	65	60	55	45	40	100	90	550	465	400	230	65.8%
Bulgaria See details	170	310	130	140	110	110	40	130	80	60	35	0	50	140	140	610	490	395	330	65.8%
Belgium See details	150	305	130	165	105	110	40	105	35	35	40	20	100	130	150	585	420	235	300	65.2%

Źródło: https://www.europeandataportal.eu/en/dashboard#2018

Serwis ocenia powyższą dojrzałość za pomocą 4 kryteriów:

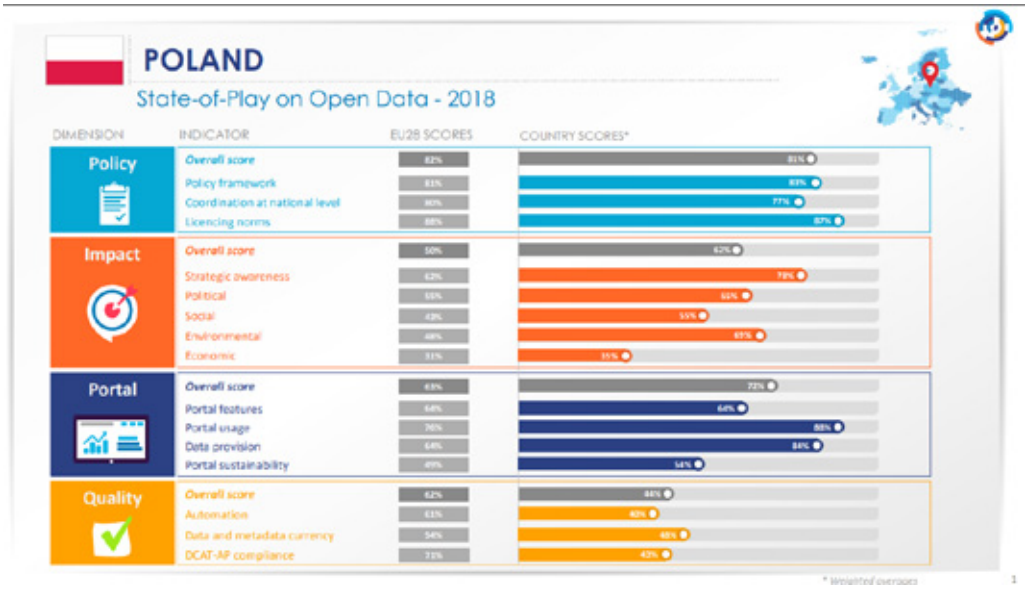
- Policy – kwestie powstania i doboru ogólnej polityki otwartości danych i standardów, koordynacji działań na poziomie krajowym.
- Portals – dostępność portali agregujących dane oraz ocena ich użytkowania, dostępnych danych.
- Impact – powszechnośćwykorzystaniaumieszczonychdanychwróżnychsektorach gospodarki i różnorodnej tematyce.
- Quality – jakość rozumiana jako zgodność z normami DCAT-AP[45], dostępność opisu udostępnionych zbiorów i automatyzacja ich udostępniania i przetwarzania.

W tych poszczególnych kryteriach oceniane jest każde państwo, szczegóły natomiast można sprawdzić w krótkich raportach opracowanych dla każdego z nich. Strukturę takiego raportu widać przykładowo dla Polski na rysunkach **3.16**, **3.17** i **3.18**.

45 https://joinup.ec.europa.eu/release/dcat-ap/12

Pierwsza strona przedstawia w szczegółach ocenę wybranego państwa w wymienionych wyżej kryteriach. Po pierwsze, znajduje się tam szczegółowa ocena, której poświęcona jest pierwsza strona (rys. **3.16**). Następnie pokazana jest zmiana pozycji w rankingu danego państwa i jej aktualna grupa związana z zaawansowaniem ekosystemu open data. Polska od roku 2017 znajduje się w grupie "Fast-trackers", którzy posiadają już portale open data. Grafika ta, przedstawiona po prawej stronie rys. **3.17**, zawiera także kluczowe informacje związane z oceną państw posiadających większy stopień zaawansowania (celem wskazania wzorców i potencjalnego wykorzystania danych pochodzących z tamtejszych rynków). Ostatnia strona raportu (rys. **3.18**) to ocena barier i najlepszych praktyk związanych z wdrażaniem i wykorzystaniem danych. Z przykładów takich stron dla poszczególnych państw można zdobyć inspirację do budowy ciekawych usług. Serwis udostępnia również szczegółowy raport w języku angielskim, który precyzuje wykorzystaną metodologię i daje ogólny wgląd w stan open data na dany rok. Raport z roku 2018 dostępny jest online[46].

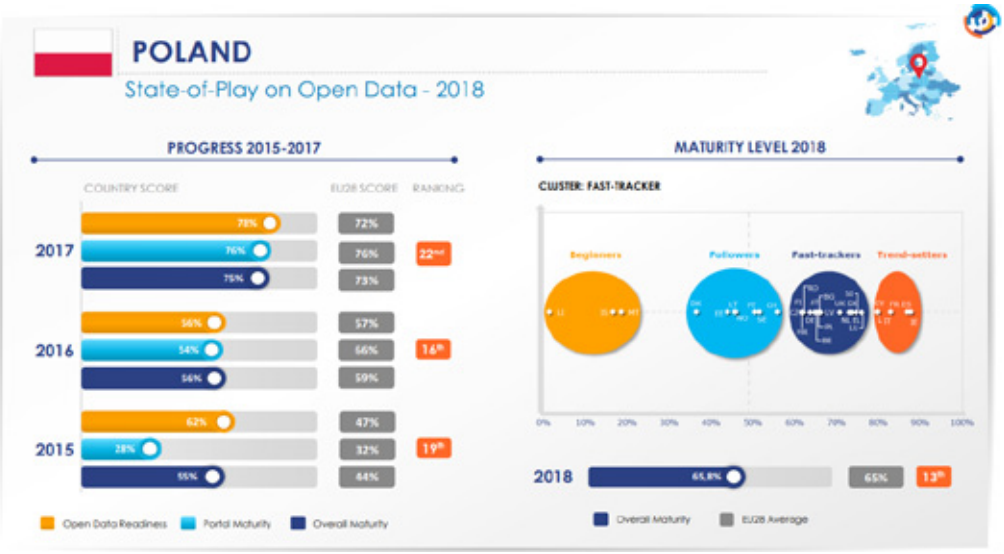
RYSUNEK 3.16. **SZCZEGÓŁOWY RAPORT OPEN DATA MATURITY DLA POLSKI – STRONA 1**



Źródło: https://www.europeandataportal.eu/sites/default/files/countryfactsheet_poland_2018.pdf

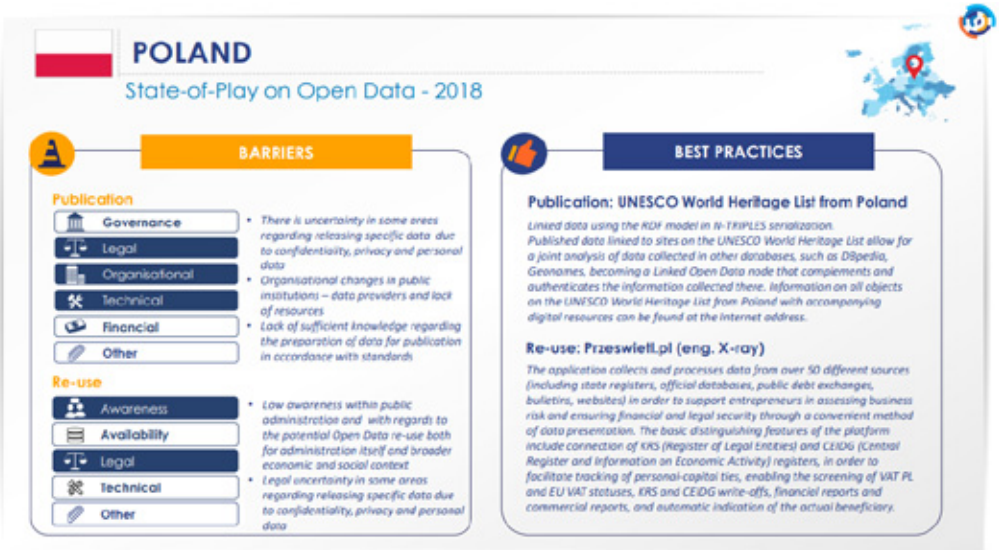
46 https://www.europeandataportal.eu/sites/default/files/edp_ landscaping_insight_report_n4_2018.pdf

RYSUNEK 3.17. **SZCZEGÓŁOWY RAPORT OPEN DATA MATURITY DLA POLSKI – STRONA 2**



Źródło: https://www.europeandataportal.eu/sites/default/files/countryfactsheet_poland_2018.pdf

RYSUNEK 3.18. **SZCZEGÓŁOWY RAPORT OPEN DATA MATURITY DLA POLSKI – STRONA 3**



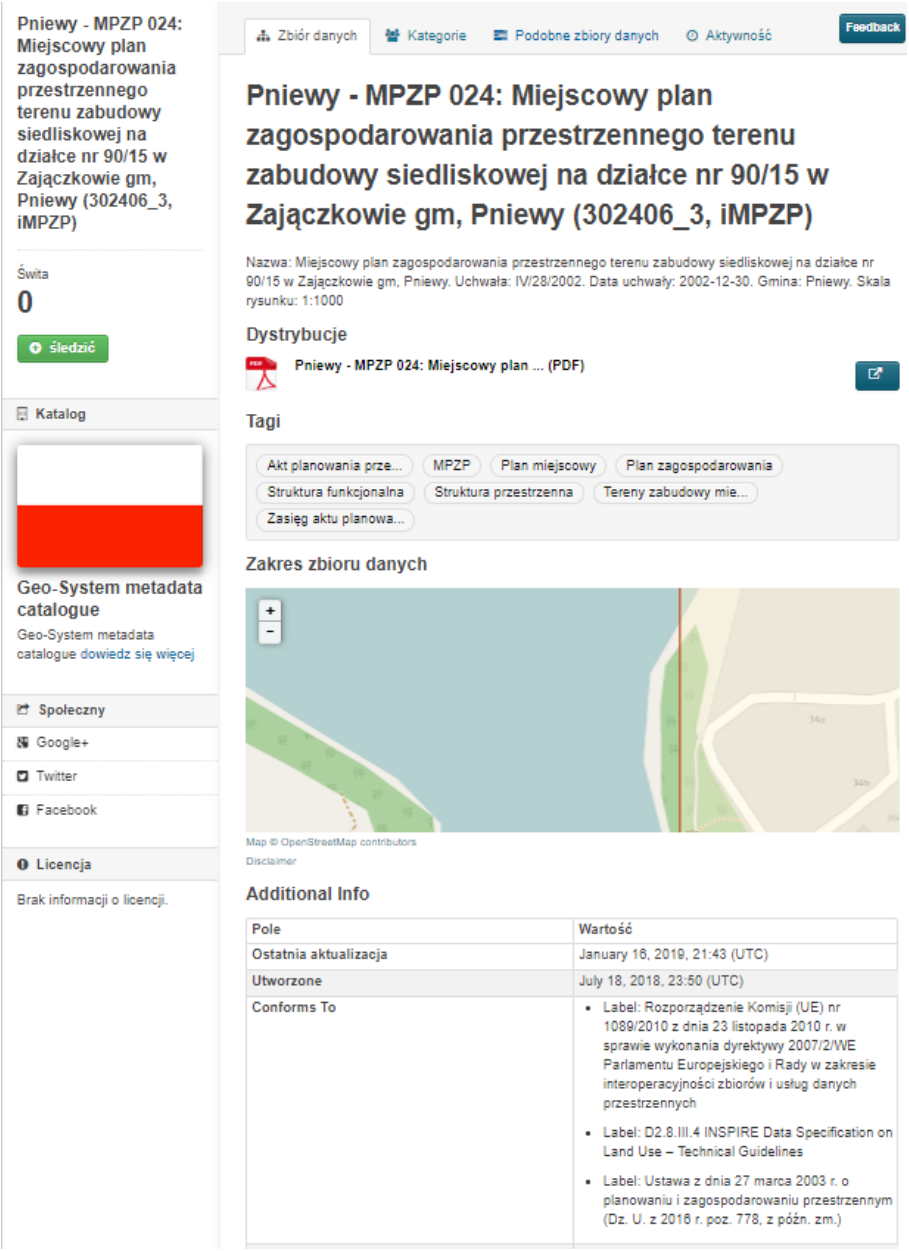
Źródło: https://www.europeandataportal.eu/sites/default/files/countryfactsheet_poland_2018.pdf

Europejski Portal Danych udostępnia również wyszukiwarkę zdeponowanych w nim zbiorów danych. Jest to wyszukiwarka oferująca tę usługę na najwyższym 5 poziomie (patrz podrozdział **2.2**). Pozwala ona na przeszukiwanie danych za pomocą zapytań w języku SPARQL. Udostępnia informacje o: kraju, formacie, licencji, kategorii tematycznej i tagach udostępnionych zbiorów danych. Serwis zawiera na moment pisania raportu ponad 888 tysięcy zbiorów danych, z czego prawie 32 tysiące dotyczące Polski. Po głębszej analizie widać jednak, iż większość zbiorów to dane o charakterze geograficznym wskazujące na zmiany w zagospodarowaniu przestrzennym i dotyczące działek, pochodzące z Geo-systemu Powiatowych Ośrodków Dokumentacji Geodezyjnej i Kartograficznej[47]. Dane udostępniane na portalu są o tyle ciekawe, że podany jest szeroki zakres metadanych wskazany w omawianym zbiorze.

Przykładowo, zbiór Pniewy – MPZP 024: Miejscowy plan zagospodarowania przestrzennego terenu zabudowy siedliskowej na działce nr 90/15 w Zajączkowie gm. Pniewy (302406_3, iMPZP)[48] posiada: tagi wskazujące na tematykę zbioru (nie jest to klasyfikacja, co pomaga w określeniu tematu), datę, powiązane dokumenty, obszar geograficzny w postaci mapy związany z zakresem dokumentu, informacje o języku udostępnionego zbioru, link i jednoznaczny identyfikator wraz z datą udostępnienia, który pozwala na dużo lepsze automatyzowanie wyszukiwania i przetwarzania udostępnionych w ten sposób zbiorów. Ten sposób katalogowania, pokazany na rysunku **3.19**, pozwala na zauważenie wszelkich zmian zaistniałych w zbiorze na przestrzeni czasu.

47 http://metadane.podgik.pl/geonetwork/srv/pol/ catalog.search/#home
48 https://www.europeandataportal.eu/data/pl/dataset/42b449727f2c-4180-ac29-2135d09ea0bb

RYSUNEK 3.19. **PRZYKŁAD UDOSTĘPNIONEGO ZBIORU DANYCH – PLATFORMA EUROPEAN DATA PORTAL**



Źródło: https://www.europeandataportal.eu/data/pl/dataset

3.3.4. Open Data for Development

Jeśli chodzi o inicjatywy społeczne (oddolne), to w zasadzie jako jedną z istotniejszych można wskazać OD4D – Open Data for Development[49]. Sama organizacja pozycjonuje się jako wiodące globalne partnerstwo, którego celem jest przyspieszenie tworzenia sterowanych lokalnie i samoutrzymujących się ekosystemów otwartych danych na całym świecie. Jej misją jest promocja lokalnego dostarczania danych otwartych wysokiej jakości. W ramach OD4D wykreowanych zostało wiele rozwiązań służących obserwacji, jak liderzy oraz innowatorzy korzystają z danych w kontaktach z rządem, społeczeństwem oraz przemysłem. Flagowe przykłady mierzenia oddziaływania danych otwartych to: mapa oddziaływania otwartych danych (Open Data Impact Map), barometr otwartych danych (Open Data Barometer) oraz indeks danych otwartych (Open Data Index).

3.3.5. Center for Open Data Enterprise

Mapa oddziaływania otwartych daych[50] jest publiczną bazą danych organizacji, które wykorzystują dane otwarte z całego świata. Jest ona rozwijana przez Center for Open Data Enterprise – CODE[51] we współpracy z OD4D. Według stanu na wrzesień 2019 w inicjatywie brało udział 1615 organizacji z 90 krajów[52]. Mapa ta jest niewątpliwie ciekawym źródeł inspiracji, jak otwarte dane mogą być wykorzystywane. Co istotne, dane o scenariuszach użycia[53] zbierane są ręcznie przez grono wolontariuszy, którzy zapewniają wysoką jakość tej bazy. Przykładowe scenariusze zostały zaprezentowane na rysunku **3.20**.

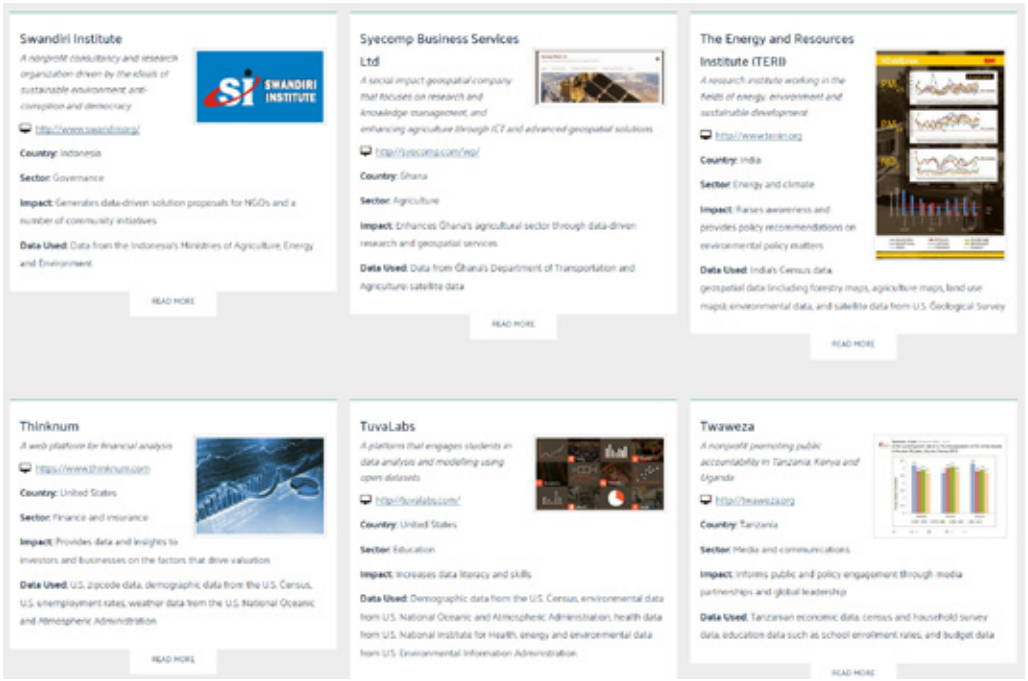
CODE nie tylko utrzymuje mapę, ale również przygotowuje okresowe raporty. Na przykład raport [**29**] prezentował wyniki z analizy 1534 przypadków użycia Open Data w 87 krajach. Na jego podstawie wyciągnięto następujące wnioski. Otwarte dane zyskują na znaczeniu jako zasób do szerszego włączania społeczeństwa do gospodarki opartej na danych. Kraje o wyższych dochodach wykazywały się średnio wyższym udziałem sektora prywatnego w wykorzystaniu danych otwartych. Sektory najczęściej wykorzystujące dane otwarte to: rząd, technologie informacyjne, badania i rozwój, konsulting. Te cztery sektory stanowią ponad połowę organizacji wykorzystujących otwarte dane, które zostały założone w latach 2006-2016. Najczęściej wykorzystywane typy danych to: działalność rządu, dane geoprzestrzenne, demograficzne i społeczne oraz pogodowe.

Raport [**29**] wskazał również cztery kategorie wykorzystania danych otwartych:

49 http://od4d.net/
50 http://opendataimpactmap.org/
51 http://www.opendataenterprise.org/index.html
52 Co ciekawe, w październiku 2017 na stronie można było znaleźć 1777 organizacji z 96 krajów.
53 http://opendataimpactmap.org/usecases.html

1. **Optymalizacja organizacji**: wzrost efektywności, lepsze poznanie rynku. Wielkie organizacje najczęściej wykorzystują dane otwarte do optymalizacjiswoich działań.
2. **Rozwój nowych produktów i usług**: dane jako zasób, uzupełnienie analityki. Nowe produkty i usługi wykorzystujące otwarte dane są najczęściej rozwijane przez firmy technologiczne, a w szczególności te zajmujące się danymi przestrzennymi.
3. **Rzecznictwo**: efektywna alokacja zasobów, lepsze tworzenie polityk. Otwarte dane poprzez odzwierciedlenie działalności instytucji rządowych mogą prowadzić do lepszej kontroli przez społeczeństwo, ale też być inspiracją do tworzenia nowych polityk uwzględniających szerszy interes społeczny.
4. **Badania**: badania przemysłowe, dziennikarstwo oparte na danych (data journalism). Ten obszar jest wyrazem dążenia do lepszego poznania otaczającej rzeczywistości z wykorzystaniem danych pochodzących z otoczenia.

RYSUNEK 3.20. **PRZYKŁADOWE SCENARIUSZE UŻYCIA OTWARTYCH DANYCH CODE**



Źródło: http://opendataimpactmap.org/usecases.html

3.3.6. Open Data Barometer

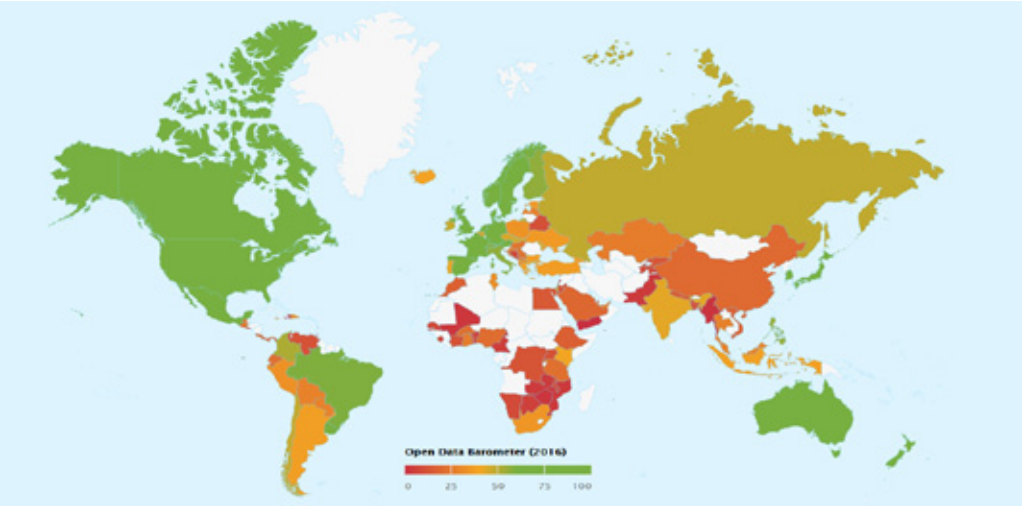
Barometr Otwartych Danych (ODB[54]) jest kolejną inicjatywą OD4D, a utrzymywany jest przez World Wide Web Foundation (WWWF[55]). Jego celem jest odkrycie wpływu działań wokół open data w różnych krajach.

Dane do ostatniej czwartej edycji opublikowanej w maju 2017 roku były zbierane od lipca 2015 do czerwca 2016, a następnie analizowane do grudnia 2016 roku. Porównano w niej 115 krajów pod względem dostarczanych danych, możliwości technicznych oraz zdefiniowanych wskaźników. Barometr szereguje kraje na podstawie następujących kryteriów:

- gotowość do zabezpieczenia korzyści wynikających z inicjatyw otwierania danych,
- implementacja programów, czyli na ile dane są faktycznie otwarte, aktualne i dostępne,
- oddziaływanie otwartych danych, czyli ocena tego, na ile otwarte dane mają rzeczywisty wpływ na analizowane dziedziny: biznes, polityka, społeczeństwo.

Bieżący ranking krajów został przedstawiony na rys. **3.21**. Dane przedstawiane na wykresach oraz mapach dostępne są online[56].

RYSUNEK 3.21. **MAPA PORÓWNUJĄCA KRAJE W OPEN DATA BAROMETER**



Źródło: https://opendatabarometer.org/4thedition/

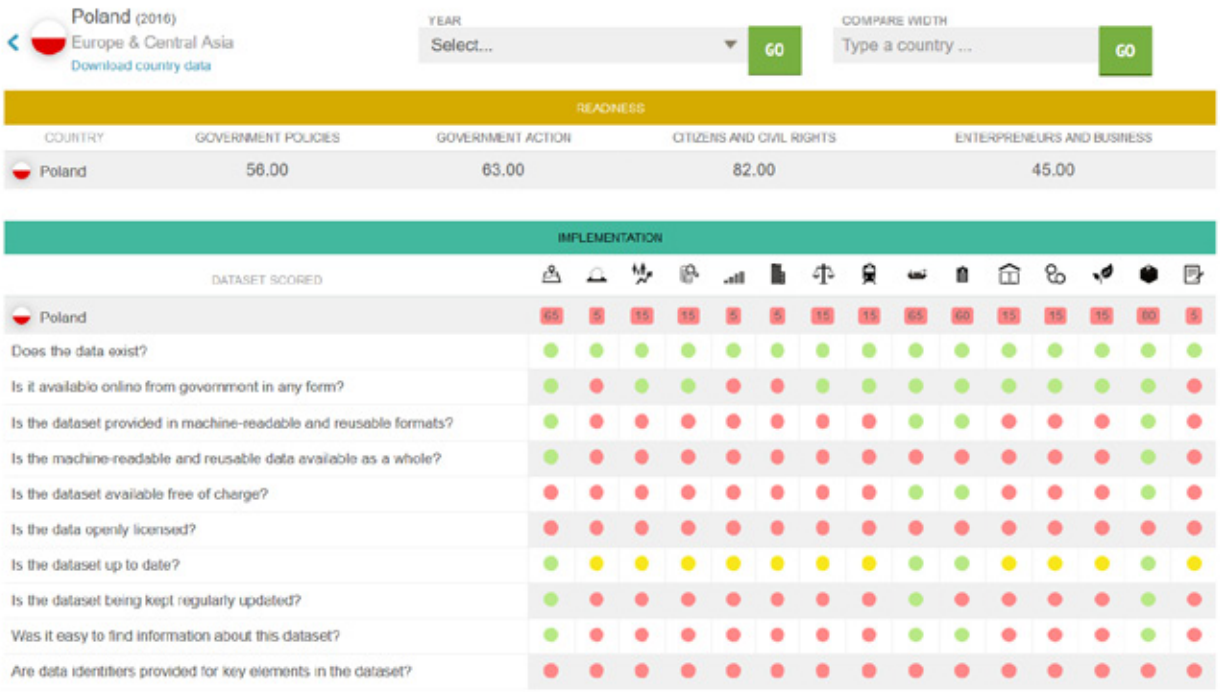
54 http://opendatabarometer.org/barometer/
55 https://webfoundation.org/
56 https://opendatabarometer.org/4thedition/data/

Wybór kraju z mapy powoduje przejście do szczegółowej oceny kraju. Przedstawiona jest ona w postaci tzw. profilu kraju. Profil dla Polski został przedstawiony na rysunku **3.22**. Dzięki niemu widać, gdzie występują główne wyzwania. W przypadku Polski stosunkowo dobrze wyglądają zasoby mapowe. Stwarza to nadzieję na rozwój firm specjalizujących się w danych geoprzestrzennych. Powinna zastanawiać słaba otwartość istotnych danych dla działalności, takich jak własność ziemi, rejestr firm, czy informacje o zamówieniach publicznych.

Jeśli chodzi o metodykę, to szczegółowa analiza na potrzeby Barometru została przeprowadzona zgodnie z *The Open Data Charter Principles*[57]. Zgodnie z tymi zasadami dane powinny być domyślnie otwarte. Barometr dokonuje analizy dostępności i jakości 15 kluczowych zbiorów danych wśród 115 krajów. Generalnie zbiory takie w takiej czy innej formie są dostępne – 97% krajów je zbiera. Jednakże 29% tych zbiorów ciągle nie jest publikowanych i tylko 7% jest rzeczywiście danymi otwartymi. Jeśli chodzi o kryterium dostępności, to 74% analizowanych baz generalnie utrzymuje aktualność danych. Oznacza to jednak, że ciągle 1/4 zbiorów ma ograniczoną wartość. Podobnie wygląda kategoria dostępności i użyteczności – 73% zbiorów można było stosunkowo łatwo znaleźć. Około 10% wszystkich analizowanych zbiorów nie było dostępnych za darmo. W kategorii porównywalności i współpracy sytuacja wygląda gorzej – nieco ponad połowa danych (53%) jest dostępna w formacie pozwalającym na wykorzystanie w przetwarzaniu maszynowym. Niestety z tych danych tylko 24% zbiorów może być pobrana hurtowo (zrzut zamiast crawlingu lub API). Zastanawiano się również, czy otwieranie danych służy zwiększeniu zaangażowania obywateli. Wpływ ten na razie jest bardzo słaby – oceniany jest na 1,2 pkt. na 10 możliwych. Średnio kraje radzą sobie lepiej z zaangażowaniem rządów w społeczności obywatelskie – ocena 4,23 na 10 możliwych. Ostatnia kategoria to rozwój i innowacje. Tutaj tylko 6% rządów może pochwalić się oddziaływaniem na włączanie marginalizowanych społeczności do korzystania z nurtu open data. Jeśli chodzi o dostępność danych służących innowacjom (np. transport publiczny), to tylko 8% zbiorów danych jest rzeczywiście otwartych. W tej ostatniej kategorii w istocie występują dane, które mogą być najbardziej interesujące ze względu na swoją unikatowość – wychodzą poza przyjęty schemat rządowy.

57 http://opendatacharter.net/

RYSUNEK 3.22. **PROFIL POLSKI DOSTĘPY W OPEN BAROMETER**



Źródło: https://opendatabarometer.org/

3.3.7. Open Data Index

Podobną rolę do Barometru pełni Indeks Otwartych Danych (Open Data Index – ODI[58]). Zmierza on do przygotowywania rankingu krajów w oparciu o ich dojrzałość rozwoju otwartych danych. Jest utrzymaniem zajmuje się fundacja Open Knowledge[59].

Oba projekty – Barometr oraz Indeks – różnią się pod względem podejścia metodycznego oraz wykorzystanych danych. Indeks bada stan otwartości rządu poprzez zbierane wśród społeczności ankiety oceniające otwartość zbiorów – skupia się zatem bardziej na perspektywie obywatela.

Open Data Indeks mierzy stan otwartych danych na podstawie dobrowolnych badań ankietowych, których celem jest określenie i ocena poziomu otwartości (w rozumieniu open data)

58 http://index.okfn.org/
59 https://okfn.org/

udostępnionych zbiorów rządowych. Zbiory danych przypisane są do jednej z 15 kategorii[60]. We wrześniu 2019 dostępnych było 166 otwartych zbiorów danych uwidocznionych w 1410 wpisach oraz 94 przypisane do określonych lokalizacji.

Jeśli chodzi o wynik Polski w obu tych rankingach, to Barometr (ODB) pokazuje 47. miejsce z wynikiem 34/100,[61] natomiast Indeks (ODI) wskazuje 28. miejsce z wynikiem 49%.

3.3.8. Inicjatywy polskie i lokalne

Wśród lokalnych inicjatyw wskazać należy przede wszystkim konkursy mające na celu opracowanie rozwiązań programistycznych (usług, aplikacji) i gotowych produktów bazujących na otwartych danych. Jedną z takich inicjatyw był hackathon BIHAPI (Business Intelligence Hackathon API) organizowany ostatnio w 2015 przez Orange we współpracy z polskimi miastami oraz przy wsparciu merytorycznym Uniwersytetu Ekonomicznego w Poznaniu, zachęcający do korzystania z zasobów otwartych danych tychże miast. Najciekawszym elementem serwisu z perspektywy tego raportu jest lista API opublikowana dla każdego z miast będących partnerami konkursu m.in. Warszawę, Gdańsk, Poznań, Kraków i województwo małopolskie – przykład usług można zobaczyć na rys. **3.23**. Zakres ten w większości pokrywa się z udostępnionymi na dzień dzisiejszy API, jednak podaje ten katalog w dużo wygodniejszy sposób. Dzięki niemu można się zorientować, jakie usługi mogą być dostępne w poszczególnych miastach.

Także administracja rządowa w Polsce zachęca do wypracowywania rozwiązań w oparciu o open data. Ministerstwo Cyfryzacji organizowało od 2017 r. hackathony[62] mające na celu popularyzację budowania aplikacji w oparciu o otwarte zbiory danych [**38**, **39**]. Swój hackathon miał również w 2018 roku Główny Urząd Statystyczny [**62**]. Przegląd wypracowanych rozwiązań (z którymi można się wstępnie zapoznać przeglądając podane źródła) wskazuje na możliwości zastosowania już udostępnionych danych i podaje ciekawe przykłady pozwalające na wskazanie obszarów, w których rozwiązania mogą przyczynić się m.in. do zwiększania jakości usług publicznych. Pewnym problemem pozostaje wdrażanie wypracowanych aplikacji i usług. Często wiążą się one z procedurą przetargową, co opóźnia i często uniemożliwia wdrożenia. Warto też wskazać na zachęcanie do korzystania podczas konkursów z usług, które są normalnie dostępne, a nie udostępniania zbiorów jedynie w celu przeprowadzenia konkursu. Trzeba jednak zaznaczyć, że w ciągu ostatnich lat nastąpiła w tych obszarach znacząca poprawa. W związku jednak z powyższymi problemami połączenie projektu z wdrażaniem jednego lub kilku zwycięskich rozwiązań może wydawać się rozwiązaniem dużo bardziej praktycznym. Jest to identyczna rekomendacja, jak dla podmiotów gospodarczych, wskazana w podrozdziale **5.2** opisującym tworzenie konkursów innowacyjności.

60 https://index.okfn.org/methodology/
61 Nastąpił spadek o 14 pozycji w stosunku do poprzedniej edycji.
62 http://hackathon.gov.pl/

RYSUNEK 3.23. **FRAGMENT STRONY Z LISTĄ API DOSTĘPNYCH DLA WARSZAWY W KONKURSIE BIHAPI**

Źródło: http://www.bihapi.pl/

3.4. Zasoby referencyjne

Wśród zbiorów danych udostępnionych przez różnorodne instytucje wskazać można także zbiory i bazy danych prowadzone lokalnie (jedynie w Polsce) i globalnie przez różnego rodzaju instytucje. Do tych drugich możemy zaliczyć opracowania udostępniane przez urzędy statystyczne, klasyfikacje towarów i przedsiębiorstw czy usługi o charakterze geograficznym.

3.4.1. Dane statystyczne

Główny Urząd Statystyczny – GUS

Mówiąc o danych statystycznych w przypadku Polski obraz udostępnionej infrastruktury wygląda bardzo korzystnie, ponieważ uruchomiony został serwis API[63] prowadzony przez Główny Urząd Statystyczny, umożliwiający dostęp przez API do systemów:

- REGON,
- TERYT,
- BDL (Bank Danych Lokalnych).

Na głównej stronie projektu (widocznej na rysunku **3.24**) zobaczyć można statystyki wykorzystania usług w interaktywnej formie. Wskazują one na znaczące wykorzystanie udostępnionych interfejsów w ostatnim czasie (w momencie pisania tego raportu).

Najważniejszą usługą z perspektywy zastosowania dostępnych danych jest Bank Danych Lokalnych. Usługa ta umożliwia dostęp do szerokiego zbioru danych statystycznych w podziale na jednostki terytorialne. Serwis online[64] pozwala na przeglądanie danych według dziedzin i jednostek terytorialnych. Wybór statystyki według dziedziny pozwala po określeniu interesującego nas wskaźnika na pokazanie go dla interesujących nas jednostek geograficznych. Wyświetlenie danych składa się z 3 etapów: wyboru dziedziny danych, wyboru odpowiednich statystyk (lewa strona rys. **3.25**) i na koniec wyboru jednostek terytorialnych, dla których mają być zaprezentowane wybrane statystyki (prawa strona rys. **3.25**). Dostępna jest także procedura pozwalająca na wybór jednostki terytorialnej w pierwszej kolejności. Następnie zaznaczone dane są prezentowane w tabeli (rys. **3.26**), a uzyskane dane można pobrać w formacie CSV lub XSLX (format programu Excel).

Dane zwracane przez interfejs BDL mogą przyjąć formę: tablicy wielowymiarowej lub tablicy relacyjnej. Nie jest to standardowy format, w którym przygotowywana jest większość plików

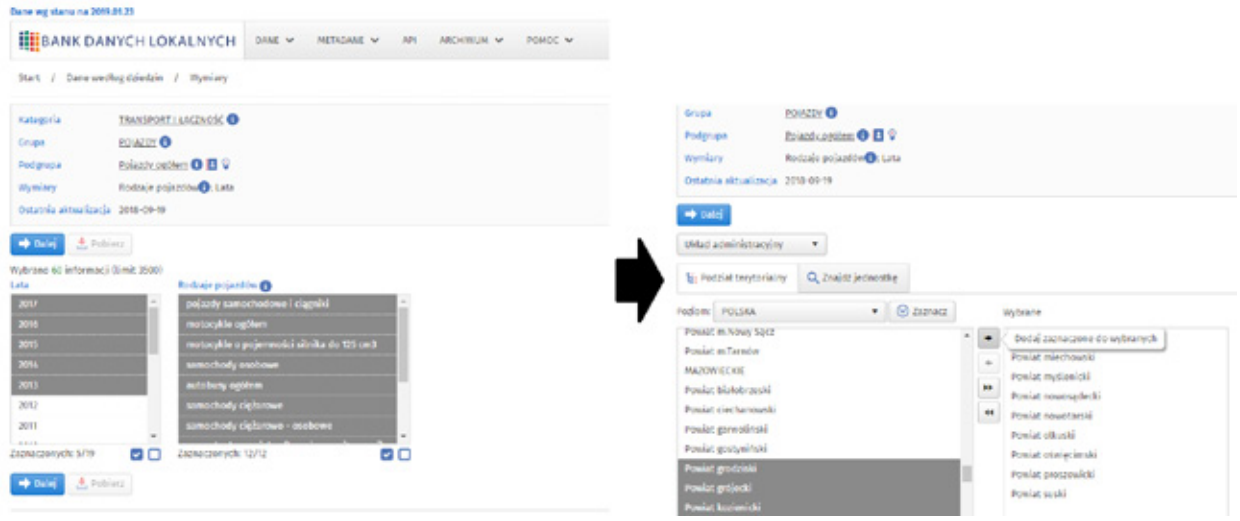
63 https://api.stat.gov.pl/
64 https://bdl.stat.gov.pl/BDL/start

z danymi, gdzie mamy do czynienia z bardzo wieloma kolumnami. Dane w formacie **tablicy relacyjnej** zawierają wszystkie informacje potrzebne do odczytania statystyki w jednym wierszu – każdy wiersz jest informacją o jakiejś zmiennej z przypisanymi odpowiednimi kategoriami. Przykład danych zwróconych w takim formacie dla jednej zmiennej i okresu 4 lat pobrany dla dwóch jednostek terytorialnych jest widoczny w listingu**3.7**. Te same dane mogą być zwrócone w formacie **tablicy wielowymiarowej**, która odpowiada bardziej standardowemu formatowi danych, gdzie mamy do czynienia z wieloma kolumnami. Dla każdego okresu konstruowana jest osobna kolumna wskazująca na wartości dla danej zmiennej. Każdy wiersz zawiera każdy okres dla poszczególnej zmiennej w osobnej kolumnie, np. *"motorowery-;2016;[szt.]"*, co można zauważyć na podstawie formatu pokazanego w listingu **3.8**.

RYSUNEK 3.24. **STRONA INTERNETOWA GUS PRZEDSTAWIAJĄCA NAJNOWSZE INTERFEJSY API**

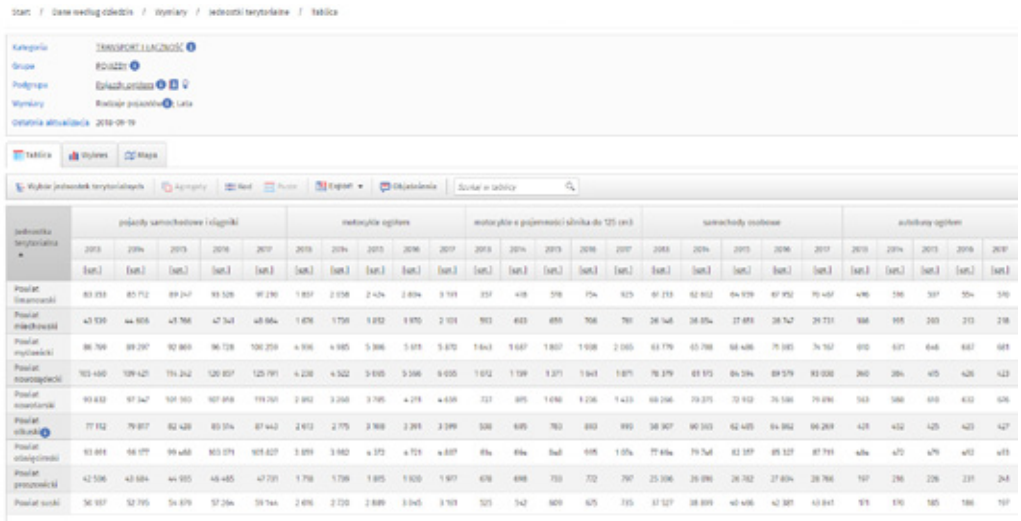
Źródło: https://api.stat.gov.pl/

RYSUNEK 3.25. **SCHEMAT POBIERANIA DANYCH WEDŁUG DZIEDZINY W SERWISIE GUS BANK DANYCH LOKALNYCH**



Źródło: https://bdl.stat.gov.pl/BDL/start

RYSUNEK 3.26. **WYNIK WYŚWIETLONEGO ZBIORU DANYCH Z SERWISU BDL**



Źródło: https://bdl.stat.gov.pl/BDL/start

KOD ŹRÓDŁOWY 3.7. **PRZYKŁAD DANYCH POBRANYCH Z BDL W FORMACIE TABLICY RELACYJNEJ CSV**

```
1  "Kod";"Nazwa";"Rodzaje pojazdów";"Rok";"Wartosc";"Jednostka miary";"Atrybut";
2  0202000;"Powiat dzierżoniowski";"motorowery";"2014";2260;"szt.";"";
3  0202000;"Powiat dzierżoniowski";"motorowery";"2015";2383;"szt.";"";
4  0202000;"Powiat dzierżoniowski";"motorowery";"2016";2463;"szt.";"";
5  0202000;"Powiat dzierżoniowski";"motorowery";"2017";2546;"szt.";"";
6  0203000;"Powiat głogowski";"motorowery";"2014";2334;"szt.";"";
7  0203000;"Powiat głogowski";"motorowery";"2015";2378;"szt.";"";
8  0203000;"Powiat głogowski";"motorowery";"2016";2400;"szt.";"";
9  0203000;"Powiat głogowski";"motorowery";"2017";2416;"szt.";"";
```

KOD ŹRÓDŁOWY 3.8. **PRZYKŁAD DANYCH POBRANYCH Z BDL W FORMACIE TABLICY WIELOWYMIAROWEJ CSV**

```
1  "Kod";"Nazwa";"motorowery;2015;[szt.]";"motorowery;2016;[szt.]";"motorowery;2017;[szt.]";
2  0202000;"Powiat dzierżoniowski";2260;2383;2463;2546;
3  0203000;"Powiat głogowski";2334;2378;2400;2416;
```

Skorzystanie z API BDL wymaga jedynie znajomości adresu URL, jednak wprowadzone są limity dla liczby zapytań. Możliwe jest ich zwiększenie po wcześniejszej rejestracji. Do rejestracji wymagany jest jedynie adres mailowy, po podaniu którego otrzymuje się unikalny klucz. Krótko charakteryzując działanie API, przegląd możliwych do uzyskania danych jest zaprezentowany w zakładce "Instrukcja" na stronie samego BDL[65].

Przykładowo, chcemy odnaleźć powyższe dane (listing **3.8**) za pomocą API.

1. Najpierw za pomocą zapytania szukamy zmiennej wskazującej na liczbę motorowerów[66].
2. Po uzyskaniu z API informacji (listing **3.9**) uzyskujemy ID potrzebnej nam zmiennej – 32560.
3. Kontynuując powyższą logiką wyszukujemy jednostki (powiaty nas interesujące) korzystając z funkcji API[67] i https://bdl.stat.gov.pl/api/v1/units/search?name= dzier%C5%BConiowski&format=json.
4. Uzyskujemy ID systemuTERYT: 030210302000 i 030210203000 – te same dane na niższym poziomie znajdowały się w danych pobranych za pomocą serwisu online.
5. Znając już zmienne nas interesujące, możemy skompletować zapytanie je zwracające. Przykładowo – możemy zwrócić wyniki dla dwóch interesujących nas jednostek[68]. Otrzymany wynik można zaobserwować w listingu **3.10**.

65 https://api.stat.gov.pl/Home/BdlApi
66 https://bdl.stat.gov.pl/api/v1/variables/search?name= motorowery&format=json
67 https://bdl.stat.gov.pl/api/v1/units/search?name= g%C5%82ogowski&format=json
68 https://bdl.stat.gov.pl/api/v1/data/by-unit/030210302000?var-id= 32560&format=json&year=2014&year=2015&year=2016&year=2017 i https: //bdl.stat.gov.pl/api/v1/data/by-unit/030210203000?var-id=32560&format= json&year=2014&year=2015&year=2016&year=2017

KOD ŹRÓDŁOWY 3.9. **PRZYKŁAD IDENTYFIKATORA ZMIENNEJ POBRANY Z BDL API**

```
{"id":32560,
    "subjectId":"P1733",
    "n1":"motorowery",
    "level":5,
    "measureUnitId":8,
    "measureUnitName":"szt."
}
```

KOD ŹRÓDŁOWY 3.10. **PRZYKŁAD DANYCH POBRANYCH Z BDL API W FORMACIE JSON**

```
{"totalRecords":1,"unitId":"030210302000",
    "unitName":"Powiat dzierżoniowski",
    "aggregateId":1,
    "results":
    [{"id":32560,
        "measureUnitId":8,
        "lastUpdate":"2018-09-19T11:28:05.903",
        "values":[
            {"year":"2014","val":2260.0,"attrId":1},
            {"year":"2015","val":2383.0,"attrId":1},
            {"year":"2016","val":2463.0,"attrId":1},
            {"year":"2017","val":2546.0,"attrId":1}
            ]
        }]
}
```

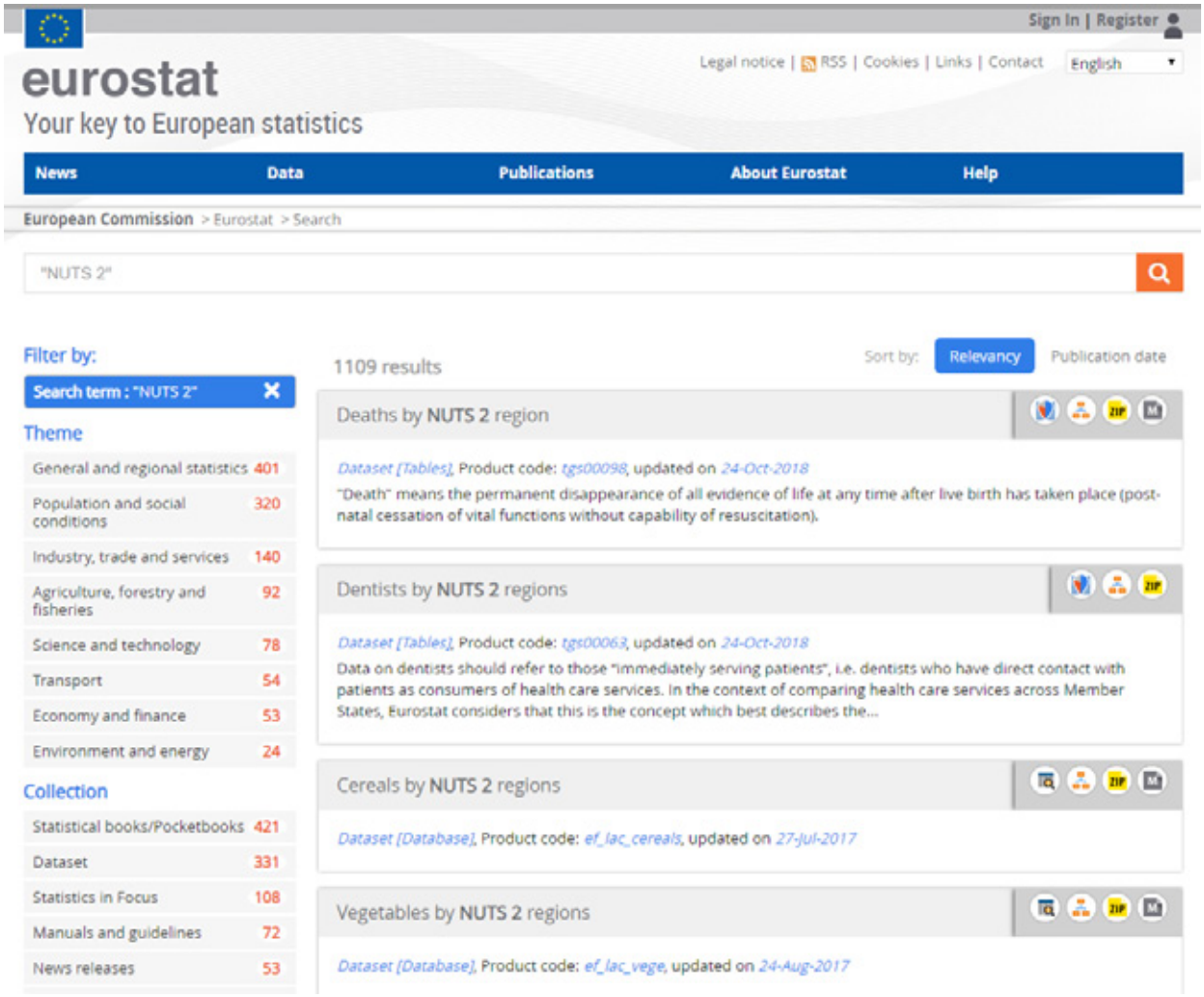
Eurostat

Innym zbiorem danych udostępniającym dane o charakterze statystycznym jest Eurostat. Pozwala on na przeglądanie danych statystycznych w wielu udostępnionych kategoriach[69]. Interfejs graficzny pozwala na wyszukanie odpowiednich danych w podziale na kategorie lub wyszukiwanie interesującego nas zbioru poprzez skorzystanie z wyszukiwarki w prawym górnym rogu (patrz rys. **3.27**). System działa podobnie jak polski serwis BDL. Po nawigacji do interesującego nas zbioru skorzystać można z aplikacji pozwalającej na wybranie interesujących nas zmiennych i obszarów (rys. **3.28**) oraz pobranie danych w wielu formatach m.in: XLSX, CSV, PDF.

Serwis EUROSTAT pozwala na przeszukiwanie danych zgodnie z **rejonami w podziale NUTS**[70] do poziomu 3. Pozwala to na wyszukiwanie danych w podziale na województwa (NUTS 2) oraz podregiony (zgrupowania powiatów) na poziomie 3. Serwis posiada również interfejs API, udostępniony opis adresu URL, który opisuje korzystanie z niego można zauważyć na rys. **3.29**.

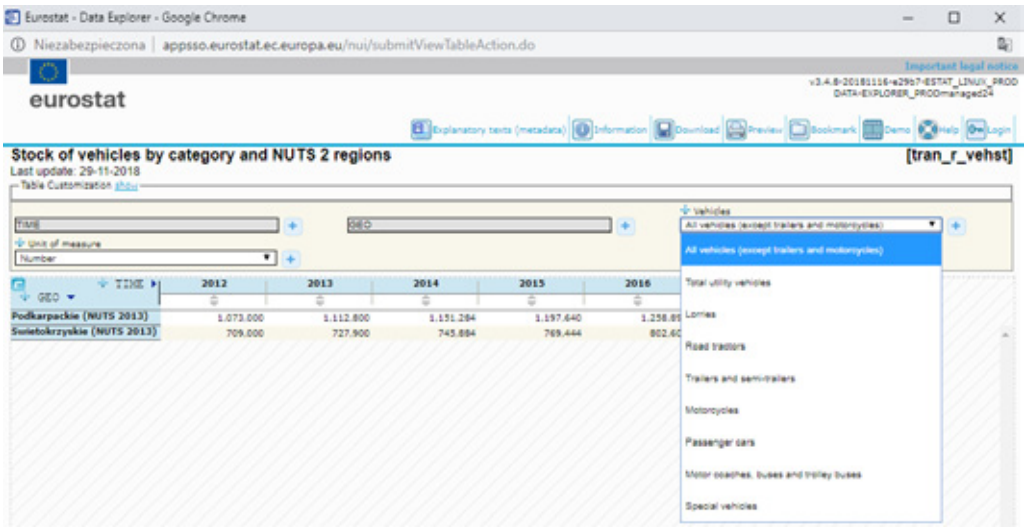
69 https://ec.europa.eu/eurostat/data/database
70 https://pl.wikipedia.org/wiki/NUTS

RYSUNEK 3.27. **STRONA GŁÓWNA WYSZUKIWANIA DANYCH SERWISU EUROSTAT**



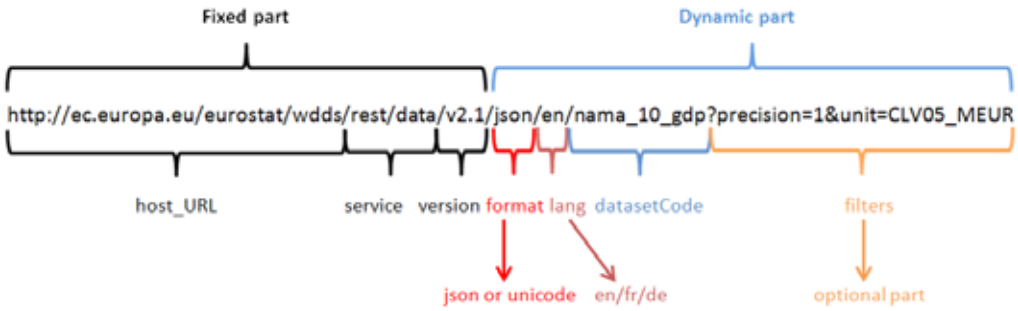
Źródło: https://ec.europa.eu/eurostat/data/database

RYSUNEK 3.28. **INTERAKTYWNY SERWIS EUROSTAT PRZEDSTAWIAJĄCY WYBRANY ZBIÓR DANYCH DLA DWÓCH OBSZARÓW NUTS 2 DLA POLSKI**



Źródło: https://ec.europa.eu/eurostat/data/database

RYSUNEK 3.29. **FORMAT API UDOSTĘPNIANY PRZEZ SERWIS EUROSTAT**

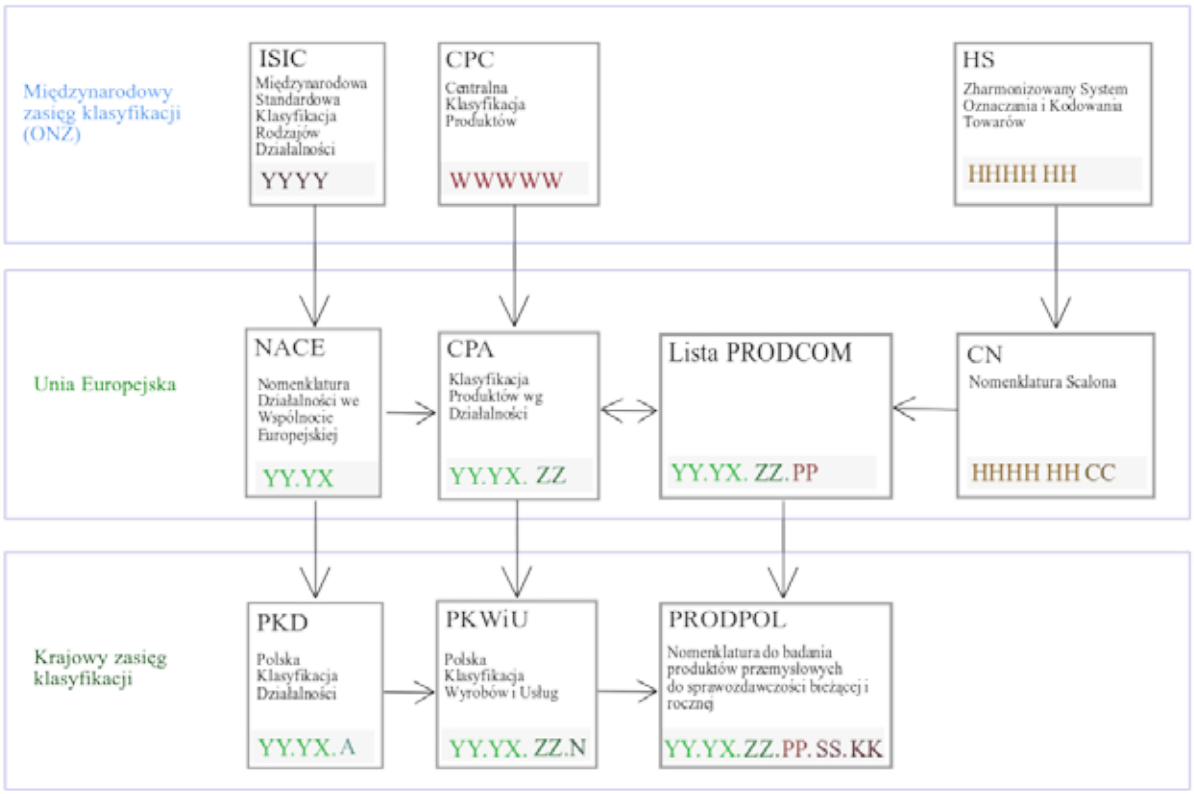


Źródło: https://ec.europa.eu/eurostat/web/json-and-unicode-web-services/ getting-started/rest-request

3.4.2. Klasyfikacje przemysłowe

W celu przetwarzania informacji o przedsiębiorstwach przydatna może być ich klasyfikacja w podziale na rodzaj działalności. Istnieje wiele różnych powszechnie stosowanych klasyfikacji (co widać na rys. **3.30**), jednak najważniejsza z perspektywy Polski jest klasyfikacja PKD.

RYSUNEK 3.30. **POWIĄZANIA KLASYFIKACJI DZIAŁALNOŚCI, PRODUKTÓW, USŁUG I TOWARÓW**



Źródło: https://upload.wikimedia.org/wikipedia/commons/1/18/Powiazania_ klasyfikacji.svg

PKD, czyli **polska klasyfikacja działalności**[71]**,** jest umownie przyjętym podziałem rodzajów działalności prowadzonej przez podmioty gospodarcze. Klasyfikacja ta ulegała zmianie na przestrzeni lat – zgodnie z rozporządzeniami Rady Ministrów. Najnowsza klasyfikacja to PKD 2007. Kod przypisywany każdej z działalności składa się z pięciu poziomów:

71 https://pl.wikipedia.org/ wiki/Polska_Klasyfikacja_Dzia%C5%82alno%C5%9Bci

1. Sekcja – poziom pierwszy, wyrażana jedną literą (np. Sekcja K – Działalność finansowa i ubezpieczeniowa), dzieli działalność na 21 grup.
2. Dział – poziom drugi, wyrażany dwucyfrowym kodem numerycznym (np. 66 Działalność wspomagająca usługi finansowe oraz ubezpieczenia i fundusze emerytalne).
3. Grupa – poziom trzeci, (np. 64.1 Pośrednictwo pieniężne) oznaczony trzycyfrowym kodem, grupujący działalności gdzie można wyodrębnić dalszą specjalizację działalności.
4. Klasa – poziom czwarty (np. 64.19 – Pozostałe pośrednictwo pieniężne), oznaczona czterocyfrowym kodem numerycznym. Zgodna z klasyfikacją NACE Rev.2
5. Podklasa – poziom piąty (np. 64.19.Z) oznaczony pięcioznakowym kodem alfanumerycznym. Powstały w celu wyodrębnienia różnić pomiędzy specyfiką polskiej gospodarki a klasyfikacjami międzynarodowymi. Jeżeli nie wprowadzano żadnych zmian, wprowadzona jest litera Z na końcu kodu.

Informacje o klasyfikacji PKD i innych klasyfikacjach można także odnaleźć na stronie GUS[72].

3.5. Dane geograficzne

3.5.1. Polskie dane geoprzestrzenne

Polska posiada bardzo rozbudowany zestaw różnych systemów danych geoprzestrzennych, m.in. map z Geoportalu Infrastruktury Informacji Przestrzennej[73]. Szeroki zestaw usług o charakterze geograficznym wymusił standaryzację formatów wykorzystywanych w celu przetwarzania informacji geograficznych. Najczęściej używanym sposobem dostarczania usług publicznych o charakterze geograficznym w Polsce jest Web Map Service (WMS), pozwalający na udostępnianie map za pomocą interfejsu HTTP. Jest to kompleksowe rozwiązanie zakładające działanie serwera, który na podstawie żądanych parametrów przesyła wygenerowany obraz mapy (JPG, PNG), korzystając z geograficznych baz danych (np. PostGIS lub standardowych baz SQL pozwalających na zapis danych przestrzennych) i plików z formatem pozwalającym na zapis informacji geograficznych (np. Shapefile, GeoJSON, GML).

Systemy utrzymywane przez administrację w Polsce tworzą Krajową Infrastrukturę Informacji Przestrzennej (KIIP). Udostępnia ona zasoby takie jak: ortofotomapy, dane katastralne, sozologiczne, hydrologiczne, granice administracyjne oraz działki ewidencyjne [**58**, **60**]. Opierając się na wyżej wymienionych źródłach podać można usługi takie jak:

72 http://stat.gov.pl/Klasyfikacje/
73 https://geoportal.gov.pl/

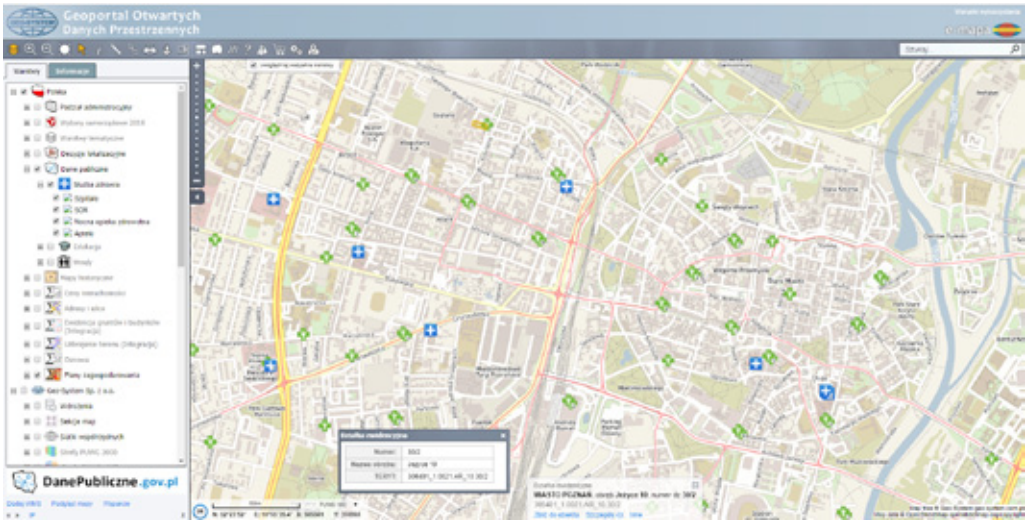
- Administracyjna Mapa Polski[74] ,
- Krajowa Integracja Ewidencji Gruntów[75],
- Ewidencja Miejscowości Ulic i Adresów[76] ,
- Dane o charakterze katastralnym[77] ,
- Ortofotomapy[78] ,
- Państwowy Rejestr Granic[79] ,
- Państwowy Rejestr Nazw Geograficznych[80] ,
- Rastrowa Mapa Hydrograficzna Polski[81] ,
- Rastrowa Mapa Sozologiczna Polski[82] ,
- Rastrowa Mapa Topograficzna Polski[83] ,
- Stacje ASG-EUPOS[84] ,
- Budynki BDO[85] ,
- Osnowa[86] ,
- Ochrona Środowiska[87] ,
- zestaw usług udostępniany przez Centrum Analiz Przestrzennych Administracji Publicznej – Główny Urząd Geodezji i Kartografii[88].

Jako że otwartych usług o charakterze geograficznym jest wiele, a powyższa lista nie jest wyczerpująca, pomocne mogą być inicjatywy mające na celu ich agregację w jednym serwisie. Jest to możliwe, ponieważ poszczególne systemy WMS dostarczają jedynie warstwy do gotowych już serwisów opartych na mapach. Jako podstawowy agregator źródeł informacji geograficznych podać można Geoportal Otwartych Danych Przestrzennych. Jest to portal mapowy o charakterze otwartym, skupiający wszystkie dostępne usługi danych przestrzennych funkcjonujące w ramach krajowej infrastruktury. Aplikacja została opracowana i jest utrzymywana

74 http://mapy.geoportal.gov.pl/wss/ service/img/guest/Administracyjna/MapServer/WMSServer
75 http://integracja.gugik.gov.pl/cgibin/KrajowaIntegracjaEwidencjiGruntow
76 http://mapy.geoportal.gov.pl/wss/service/PZGIKINSP/guest/services/G2_EMUIA_WMS/MapServer/WMSServer
77 http://sdi.geoportal.gov.pl/wms_dzkat/wmservice.aspx
78 http://sdi.geoportal.gov.pl/wms_orto/wmservice.aspx
79 http://sdi.geoportal.gov.pl/wms_prg/wmservice.aspx
80 http://sdi.geoportal.gov.pl/ wms_prng/wmservice.aspx
81 http://sdi.geoportal.gov.pl/wms_hydro/wmservice.aspx
82 http://sdi.geoportal.gov.pl/wms_sozo/wmservice.aspx
83 http://sdi.geoportal.gov.pl/wms_topo/wmservice.aspx
84 http://sdi.geoportal.gov.pl/gm_wms_asg/request.aspx
85 http: //sdi.geoportal.gov.pl/wms_budynki_bdot/request.aspx
86 http://sdi.geoportal.gov.pl/WMS_OSNOWY/Request.aspx
87 http://wms.gdos.gov.pl/geoserver/wms
88 https://capap.gugik.gov.pl/cat

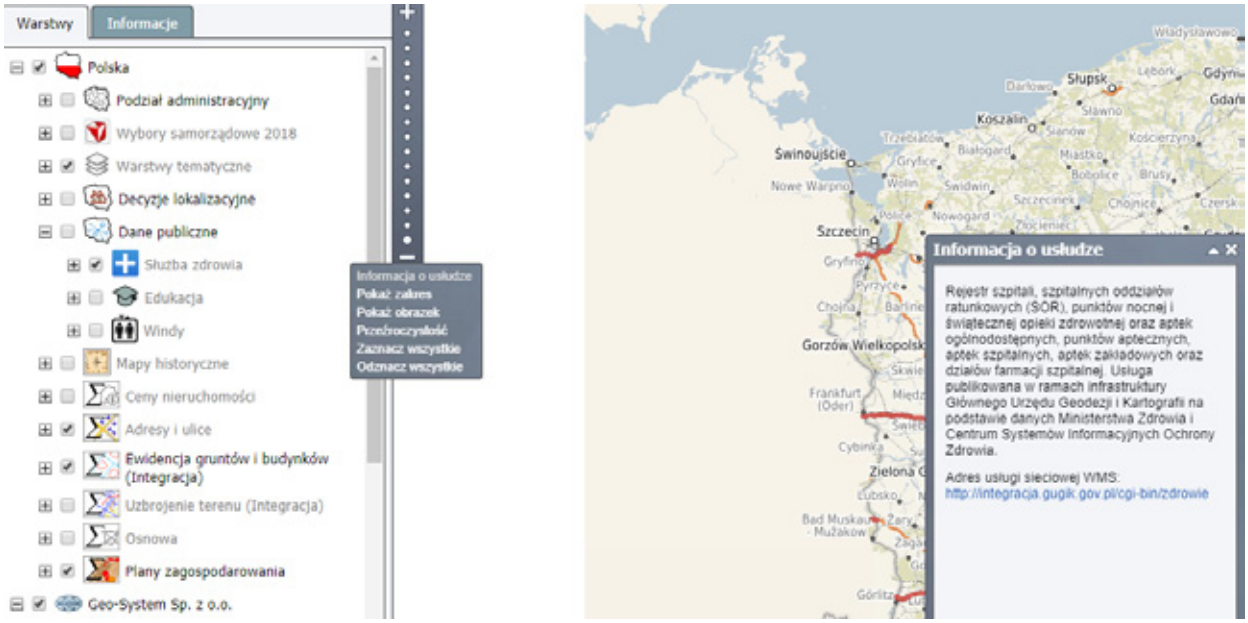
przez firmę Geo-System Sp. z o.o. z otwartych zbiorów i usług WMS o charakterze geograficznym. Sama usługa jest scenariuszem użycia dla otwartych danych geograficznych i pozwala na bardzo dokładne określenie, np. działki ewidencyjnej dla punktu, który został wybrany – rys. **3.31**. Dodatkowo usługa pozwala na uzyskanie informacji o źródle wykorzystywanej warstwy (patrz rys. **3.32**), co może być bardzo pomocne w przypadku chęci budowania usług o charakterze geograficznym z dostępnych źródeł. Jeżeli mowa o lokalnych danych dotyczących gruntów, budynków i lokali, to zgodnie z informacjami z geoportalu [**57**], dane te udostępniane w postaci usług WMS są przez starostów lub prezydentów miast na prawach powiatu. Ewidencję gruntów i budynków prowadzi się w systemie teleinformatycznym, którego podstawę stanowią komputerowe bazy danych ewidencyjnych. Dane ewidencyjne udostępniane są przez Powiatowe Ośrodki Dokumentacji Geodezyjnej i Kartograficznej (PODGiK). Dane takie są uwzględniane w Geoportalu (zarówno geoportal.gov.pl jak i Geoportalu Otwartych danych Przestrzennych) jako warstwa. Same także są udostępniane jako WMS i często indeksowane w innych usługach (przykład – rys **3.19**). Informacje te obejmują przede wszystkim: krajową Ewidencję Gruntów z informacjami o obszarze działek i budynków oraz krajową integrację Miejscowych Planów Zagospodarowania Przestrzennego. Efektem integracji powyższych źródeł jest ujednolicony dostęp do planów zabudowy widoczny jako warstwa w serwisach obu wymienionych w podrozdziale geoportali.

RYSUNEK 3.31. **MAPA SERWISU GEOPORTAL OTWARTYCH DANYCH PRZESTRZENNYCH**



Źródło: http://polska.e-mapa.net/

RYSUNEK 3.32. **PRZYKŁAD OPISU WARSTWY W SERWISIE GEOPORTALU OTWARTYCH DANYCH PRZESTRZENNYCH**



Źródło: http://polska.e-mapa.net/

Dodatkowo celem uzyskania informacji o jednostkach podziału terytorialnego (np. budując usługę agregującą obiekty zgodnie z jednostkami terytorialnymi), chcąc otrzymać listę obiektów o konkretnym charakterze czy skorzystać z wyszukiwarki adresów opartej na otwartych danych – wszystko to możliwe jest dzięki usłudze TERYT. System ten pełni rolę krajowego rejestru urzędowego podziału terytorialnego kraju. TERYT sam w sobie pozwala na indeksację i przeglądanie numerów identyfikujących poszczególne jednostki terytorialne – co zauważyć można na rys. **3.34**. System, zgodnie z oficjalną dokumentacją obejmuje podsystemy:

- TERC – identyfikatorów i nazw jednostek podziału terytorialnego (mapowanie pokazane na rys. **3.33**),
- SIMC – identyfikatorów i nazw miejscowości,
- BREC – rejonów statystycznych i obwodów spisowych,
- NOBC – identyfikacji adresowej ulic, nieruchomości, budynków i mieszkań (w ramach systemu prowadzony jest Centralny Katalog Ulic – ULIC).

Usługa jest także udostępniona w formie API przez GUS[89]. API posiada bardzo dokładną i uaktualnianą dokumentacje, dostępną w zakładce 'instrukcja', obejmującą szeroki zakres udostępnianych możliwości systemu. Dostęp wymaga rejestracji, jednak jest ona możliwa także dla konta przypisanego do osoby prywatnej.

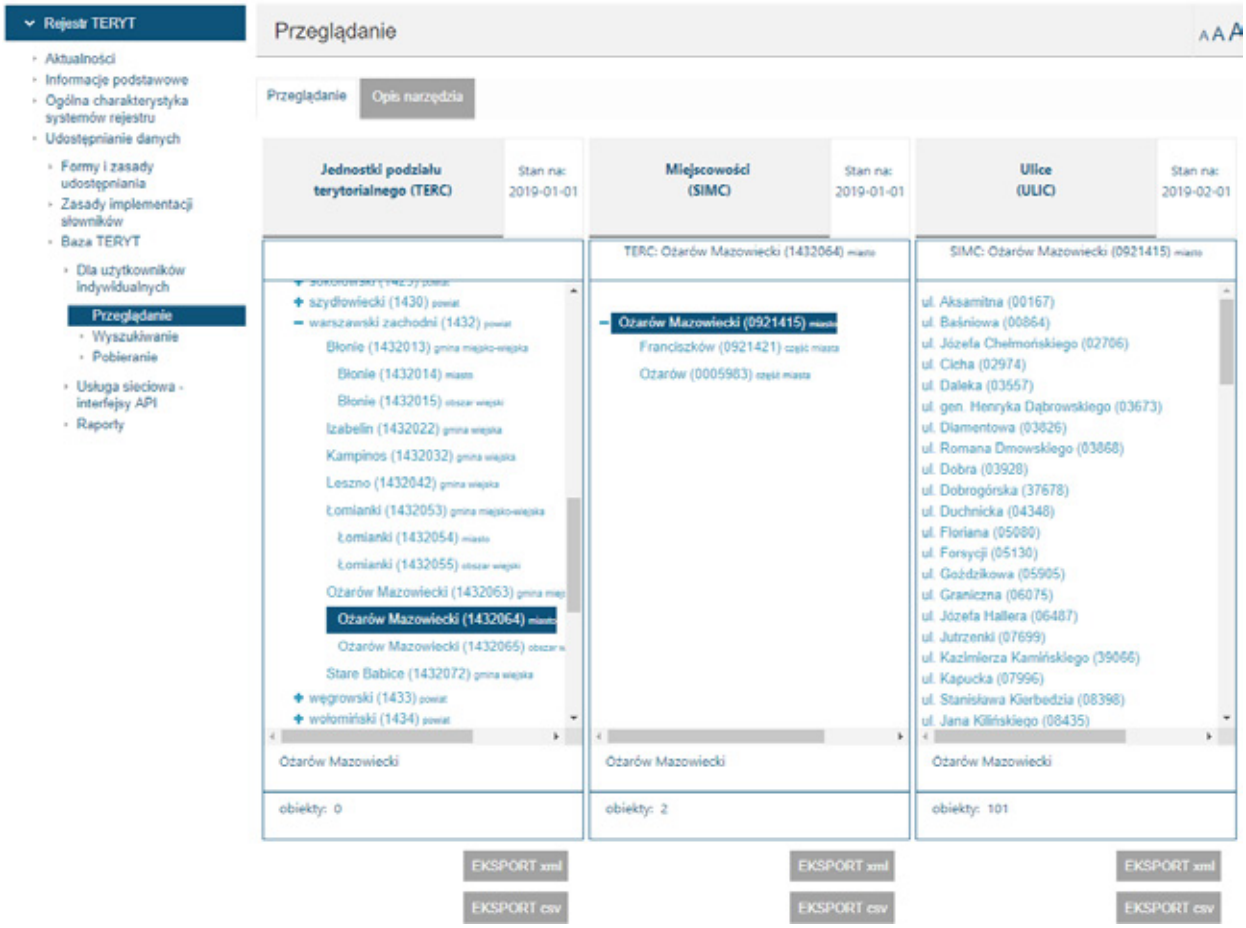
RYSUNEK 3.33. **OPIS MAPOWANIA W SYSTEMIE TERC**

CZŁON	OPIS	ZAKRES
Województwo	**Dwucyfrowy symbol województwa**, nadawany województwom ułożonym w kolejności alfabetycznej	Liczby parzyste od 02 do 98
Powiat	**Dwucyfrowy symbol powiatów**, nadawany powiatom danego województwa ułożonym w kolejności alfabetycznej, a następnie miastom na prawach powiatu	01–60 – symbol powiatu 61–99 – symbol miasta na prawach powiatu
Gmina	**Trzycyfrowy symbol gminy**: dwie pierwsze cyfry stanowią numer gminy nadawany gminom danego powiatu ułożonym alfabetycznie, począwszy od gmin miejskich, następnie wiejskich i gmin miejskich	Pierwsze dwie cyfry (symbol gminy) w przedziale od 01 do 99 Trzecia cyfra stanowi symbol rodzaju jednostki i oznacza: 1 – gmina miejska 2 – gmina wiejska 3 – gmina miejsko-wiejska 4 – miasto w gminie miejsko-wiejskiej 5 – obszar wiejski w gminie miejsko-wiejskiej 8 – dzielnice gminy Warszawa Centrum 9 – delegatury i dzielnice innych gmin miejskich

Źródło: [190]

89 https://api.stat.gov.pl/Home/TerytApi

RYSUNEK 3.34. **PRZYKŁAD Z WYSZUKIWARKI SYSTEMU TERYT**



Źródło: http://eteryt.stat.gov.pl/eTeryt/rejestr_teryt/udostepnianie_ danych/baza_teryt/uzytkownicy_indywidualni/przegladanie/ przegladanie.aspx?contrast=default

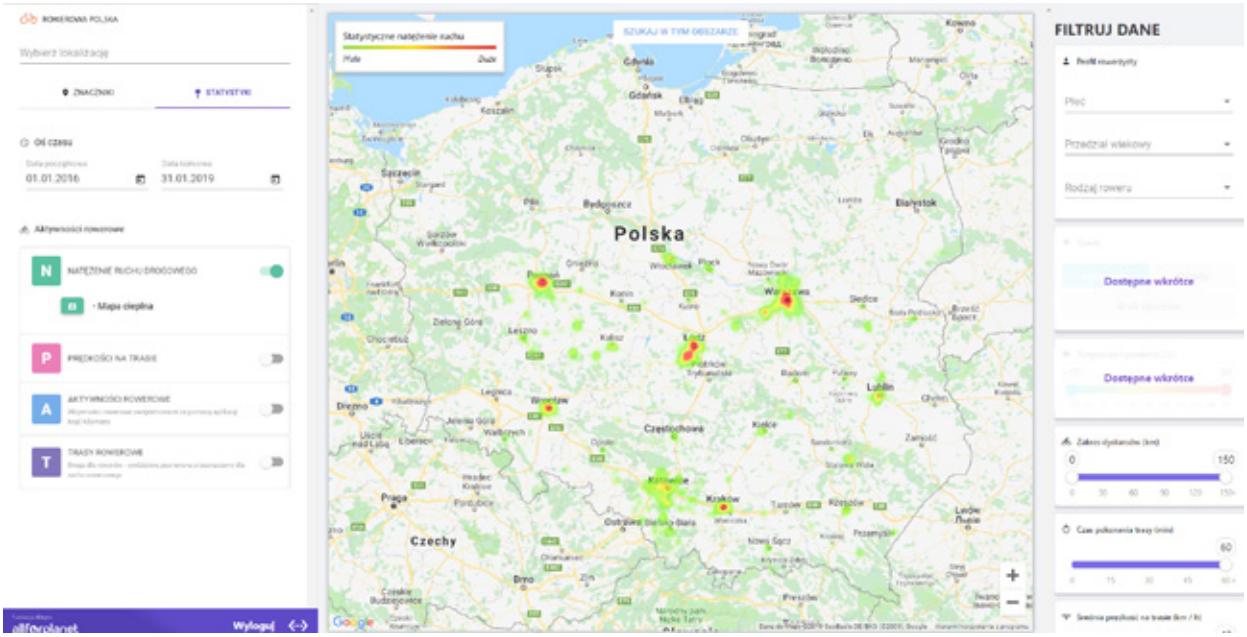
3.5.2. Dane z inicjatyw zewnętrznych

Jeżeli mowa o danych o charakterze geograficznym, wskazać można także na zbiory poświęcone popularyzacji transportu rowerowego, np. efekt zorganizowanej akcji „Kręć kilometry" – organizowanej przez fundację Allegro for planet[90]. Serwis daje dostęp do kompilacji zanonimizowanych danych zebranych w obrębie działania aplikacji. Zgodnie z oficjalną dokumentacją: API oferuje również szereg agregacji tras i danych z nimi związanych, co pozwala na spraw-

90 https://kreckilometry.pl/

niejsze budowanie wizualizacji danych. Serwis udostępnia zbierane dobrowolnie od użytkowników dane o ich trasach rowerowych wraz z ich podstawową metryką, jak i dane o ścieżkach rowerowych. Korzystanie z niego wymaga rejestracji, natomiast po jej przeprowadzeniu serwis udostępnia mapę (rys. **3.35**), pozwalającą na wyświetlenie interesujących danych na podstawie podanych parametrów, oraz serwis Rowerowa Polska API[91]. O innych zbiorach danych można przeczytać m.in. w artykule umieszczonym w serwisie „Koduj dla Polski" [**146**].

RYSUNEK 3.35. **SERWIS MAPY UDOSTĘPNIONEJ W SERWISIE ROWEROWAPOLSKA.PL**



Źródło: https://rowerowapolska.pl/

91 https://api.rowerowapolska.pl/public/api/documentation

3.5.3. Zbiory danych Smart Cities

Tematyka Smart Cities obejmuje sensory możliwe do wykorzystania w przestrzeni miejskiej jako otwarte zbiory danych. Przykładowe usługi opisane zostały w podrozdziale **3.2.3**. Tematyka poniższych usług i zbiorów ściśle związana jest z Internetem Rzeczy (Internet of Things) i inteligentnych sensorów[92], które są ciekawe z perspektywy wykorzystania dla szerokiego grona zastosowań [**12**, **106**] i podmiotów, chociażby:

- Dostawców sensorów i oprogramowania, którzy mogą na podstawie przykładów przygotować rozwiązania dla polskich miast i zbudować innowacyjne na skalę światową produkty.
- Instytucji samorządowych i spółek u znaczącym udziale państwowym, które mogą skorzystać z usług w tym zakresie, aby udostępnić usługi wysokiej jakości.
- Spółek celowych uczelni wyższych i naukowców, aby zainteresować się tą ciekawą dla administracji państwowej i samorządowej oraz przemysłu dziedziną.

Dane o takiej charakterystyce można znaleźć przeszukując rządowe i samorządowe serwisy open data, np: dla USA[93], dla UE w oparciu o opracowania i projekty naukowe [**171**] lub odwiedzając portale Open Data poszczególnych miast[94].

Można je podzielić także tematycznie, np. skupiając się na danych dotyczących rowerów[95], Smart-Grid – czyli inteligentnych czujników zużycia prądu[96], czy też urządzeń prywatnych[97]. Tematyce wykorzystania powyższych zbiorów można by poświęcić cały osobny raport, ponieważ prowadzi się badania nad ich wykorzystaniem w celu: optymalizacji transportu miejskiego, planowania przestrzennego, analizy jakości życia i wpływu miasta na środowisko, wewnętrznych analiz marketingowych i umiejscowienia nowych oddziałów dla dużych przedsiębiorstw.

Poza dynamicznymi źródłami wyróżnić można także zagregowane zbiory dostępne jako open data, przedstawione w tabeli **3.2**. W styczniu 2019 roku dostępne były dwa duże zbiory opublikowane na korzystnych licencjach, oba o charakterze badawczym. Dane takie w badaniach naukowych często koreluje się z innymi źródłami (dane z usług transportowych[98], czy dane z aktywności w social media z geotagowaniem, np. Twitter).

92 Używane w literaturze jest określenie „smart", co oznacza głównie możliwość dostępu do danych i interakcji z urządzeniem na poziomie sieci internetowej.
93 https://catalog.data.gov/dataset?tags=iot oraz https://catalog.data.gov/ dataset?tags=sensors
94 https://datosabiertos.malaga.eu/dataset
95 https://github.com/ubahnverleih/WoBike, https://data.socialbicycles.com/
96 https://data.gov.au/dataset/ds-dga-4e21dea3-9b87-4610-94c7-15a8a77907ef/details
97 http://www.social-iot.org/
98 https://github.com/fivethirtyeight/uber-tlc-foil-response, https: //www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page, https:// movement.uber.com/cities?lang=en-US

TABELA 3.2. **PRZYKŁADY DUŻYCH ZBIORÓW DANYCH O TEMATYCE SMART CITIES**

NAZWA ZBIORU	REJESTRACJA	ROZMIAR	OPIS	LICENCJA
Telecom Italia Big Data Challenge dataset[99]	Tak, na stronie wnętrznej[100]	50 GB dla każdego z miast	Zagregowane dane dotyczące komunikacji telefonicznej (2 miasta- Trento i Mediolan), udostępnione przez operatora komunikacyjnego. Uzupełnione o dane pogodowe, dot. jakości powietrza, energetyczne - Smart GRID, społecznościowe- geotagowane tweety i wiadomości z serwisów internetowych zebrane w 2014 roku. Dane są przypisane do zanonimizowanych komórek grid.	The Open Database License (ODbL)
Citypulse dataset[101]	Nie	17 GB	Dane dotyczące natężenia transportu, stanu czujników parkowania, pogody i zanieczyszczeń, różnych kategorii wydarzeń zebrane na przełomie 2014 i 2015 roku.	Creative Commons Attribution 4.0 International License

Źródło: Opracowanie własne

3.6. Publiczne dane zdrowotne

W dziedzinie danych zdrowotnych coraz bardziej widoczne są wpływy koncepcji open data. Organizacje takie jak International Medical Informatics Association[102] podejmują współpracę z podmiotami zajmującymi się inicjatywą open source w celu zapewnienia dostępności i jakości danych medycznych. Wywierają tym samym pozytywny wpływ na poziom opieki zdrowotnej na całym świecie, między innymi poprzez zmniejszanie kosztu badań.

Bogate zbiory danych udostępniane są przez instytucje rządowe. Departamenty zdrowia w różnych krajach dostrzegają konieczność publicznego dzielenia się posiadanymi danymi. Wśród najbardziej znanych przykładów serwisów tworzonych przez rządy można wymienić:

- **HealthData.gov**[103]

Projekt prowadzony przez amerykański Departament Zdrowia i Opieki Społecznej. Jego celem jest zapewnienie dostępu do rządowych danych, które mogą zostać wykorzystane z korzyścią dla zdrowia publicznego. Zbiory danych pochodzą z różnych wydziałów Departamentu Zdrowia i Opieki Społecznej, a także współpracujących z nim podmiotów, takich jak Agencja Żywności i Leków (FDA). Dotyczą one wielu zagadnień, m.in. zdrowia środowiskowego, urządzeń

99 http://theodi.fbk.eu/openbigdata/http://aris.me/contents/teaching/data-mining-2015/project/BigDataChallengeData.htm
100 https://dandelion.eu
101 http://iot.ee.surrey.ac.uk:8080/datasets.html i http://www.ict-citypulse.eu/page/content/tools-and-data-sets
102 https:// imia-medinfo.org
103 https://healthdata.gov

medycznych, zdrowia psychicznego oraz nadużywania substancji. Strona internetowa projektu posiada wyszukiwarkę, umożliwiającą przeglądanie zasobów za pomocą słów kluczowych, podziału na tematy, systemu tagów, autorów oraz licencji. Większość zbiorów dystrybuowana jest na licencji Open Data Commons Open Database Licence i innych licencjach publicznych. Niektóre z zasobów dostępne są za pomocą dedykowanych API.

- **data.gov.uk**[104]

Uruchomiony w 2009 roku projekt brytyjskiego rządu, którego misją jest udostępnianie niewrażliwych danych rządowych na zasadach open data. Zbiory podzielone są na 12 głównych kategorii, w tym zdrowie. W chwili obecnej kategoria ta zawiera 61 pozycji. Dotyczą one między innymi konsumpcji tytoniu, alkoholu i narkotyków oraz funkcjonowania szpitali. Wyszukiwarka umożliwia przeglądanie zbiorów pod kątem słów kluczowych, tematów oraz formatu danych. Większość danych publikowana jest na licencji Open Government Licence.

Dynamiczny rozwój bioinformatyki sprawia, że dostępne są również rozbudowane zasoby z zakresu genetyki. Oferowane są one m.in. przez organizacje rządowe (np. NCBI[105]) oraz uniwersytety (np. stworzona przez Uniwersytet Kalifornijski w Santa Cruz wyszukiwarka zbiorów danych Xenabrowser[106]). Inne warte wzmianki projekty w tej dziedzinie to:

- **Uniport**[107]

Inicjatywa skupiająca się na dostarczeniu wysokiej jakości danych związanych z sekwencjonowaniem białek. Oferuje swoje zasoby na licencji Creative Commons By 4.0. Podzielona jest na cztery części: *UniRef* (grupowane ciągi sekwencji), *UniParc* (archiwum sekwencji białek), *Proteomes* (zbiory proteomów – zestawów białek występujących w komórkach) i *Supporting data* (dane pomocnicze – literatura, słowa kluczowe, referencje do baz danych i inne). Większość danych udostępnianych przez Uniport pochodzi z różnych projektów sekwencjonowania ludzkiego genomu.

- **GeneOntology**[108]

Organizacja udostępniająca jeden z największych na świecie zbiorów danych na temat funkcji genów. Dane zorganizowane są w sposób umożliwiający ich odczytanie zarówno przez ludzi, jak i maszyny. Wykorzystywane są do komputerowej analizy na potrzeby biologi molekularnej i badań genetycznych.

104 https://data.gov.uk
105 https://www.ncbi.nlm.nih.gov/
106 https://xena.ucsc.edu/
107 https://www.uniprot.org/
108 http://geneontology.org/

Nowoczesne techniki analizy danych, takie jak uczenie maszynowe, mogą być wykorzystywane jako narzędzia wspierające specjalistów przy diagnozach i podejmowaniu decyzji dotyczących leczenia. Potencjał sztucznej inteligencji w tym zakresie dostrzegła między innymi firma Google, która prowadzi intensywne badania i dzieli się ich rezultatami[109]. Skupiona wokół tego problemu społeczność tworzy, publikuje i dba o dostępność różnorodnych zasobów (patrz: *Medical Data for Machine Learning*[110]). Dużą popularnością cieszą się dane wizualne – głębokie sieci neuronowe osiągają bowiem obiecujące wyniki także w analizie obrazów medycznych. Wartymi wzmianki zbiorami obrazów medycznych przeznaczonymi do uczenia maszynowego są:

- **MedPix**[111]

Darmowa baza różnego rodzaju zdjęć medycznych – obrazów rentgenowskich, mikroskopowych oraz zdjęć przypadków. Łączy zasoby wizualne i tekstowe metadane. Zawiera niemal 59 tys. zdjęć, dotyczących ponad 12 tys. pacjentów. Dane segregowane są według kilku kryteriów, takich jak miejsce występowania schorzenia, jego kategorii oraz opisów zdjęć. Możliwe jest wyszukiwanie za pomocą opisów objawów, nazw organów, opisów zdjęć i innych słów kluczowych. Użycie danych poza celami prywatnymi i edukacyjnymi wymaga uzyskania zgody od właścicieli praw do poszczególnych zasobów.

- **OASIS - Open Access Series of Imaging Studies**[112]

OSAIS to projekt, którego celem jest nieodpłatne dostarczanie środowisku naukowemu zbiorów danych związanych z neuroobrazowaniem. Najnowsza wersja zbioru (OSASIS-3) zawiera dane dotyczące 1098 pacjentów, na które składają się zapisy z niemal 2200 sesji rezonansu magnetycznego oraz ponad 1600 sesji pozytonowej tomografii emisyjnej. Jego celem jest wsparcie badań, mogących przyczynić się do rozwoju neurobiologii oraz neurochirurgii.

Należy pamiętać, iż dane medyczne są danymi szczególnego rodzaju ze względu na bardzo wysoki poziom wrażliwości. Nawet jeżeli zostały poddane anonimizacji, dostęp do nich i zasady wykorzystania mogą być bardziej skomplikowane niż w innych dziedzinach. Jako przykład można wskazać jedną z najbardziej znanych medycznych baz danych, stworzony na Massachusetts Institute of Technology projekt MIMIC[113]. Aby uzyskać do niego dostęp, należy nie tylko się zarejestrować, ale również ukończyć kurs, dotyczący obchodzenia się z wrażliwymi danymi. Przedstawienie certyfikatu ukończenia jest jednym z warunków rozpatrzenia zgłoszenia (nie

109 https://ai.google/healthcare/
110 https://github.com/beamandrew/medical-data
111 https://medpix.nlm.nih.gov
112 http://www.oasis-brains.org/
113 https://mimic.physionet.org/

gwarantuje jednak odpowiedzi pozytywnej). Podejście open data jest jednak coraz popularniejsze i można liczyć na pojawianie się w przyszłości kolejnych zasobów dostępnych publicznie i bez ograniczeń.

RYSUNEK 3.36. **PRZYKŁADY BAZ DANYCH ZDROWOTNYCH DOSTĘPNYCH ONLINE**

NO.	DATA REPOSITORY NAME	ISSUER	URL	ACCESSIBILITY	LICENSE	MACHINE READABILITY	AVAILABLE FORMATS
1	HealthData.gov	U.S. Department of Health & Human Services	https://www.healthdata.gov/	open	open	3531/3542	JSON, CSV, XML, RDF
2	DAJA.G0V.JK	United Kingdom, government digital service	https://data.gov.uk/	open	open	2121/2132	CSV, XLS, HTML
3	The human mortality database	The Human Mortality Database Project	http://www.mortality.org/	registration required	not open	12103	CSV
4	Global health observatory data	World Health Organisation	http://www.who.int/gho/database/en/	open	not open	>1000	JSON, CSV, XML, XLS
5	Big cities health coalition	A forum for the leaders of America's largest metropolitan health departments	https://bchi.bigcitieshealth.org/	open	not explicit	53/53	CSV
6	DATA GO JP	Cabinet secretariat of Japan	http://www.data.go.jp	open	open	112/624	CLS, HTML, PDF
7	Dryad	Non-profit organisation	http://datadryad.org	open	open	974/974	XLS, MATLAB, RMD, SOLR
8	UKDA	Academicgroup	http://www.data-archive.ac.uk	open	open	>1502	PDF, XLSX, CSV
9	Physionet	NIH granted project	http://www.physionet.org	open butpartial registration required	open`	>111	R, MATLAB, Database, API
10	Open Health Data dataverse	Harvard university	https://dataverse.harvard.edu/dataverse/openhealthdata	open	not open	About 62900	Text, CSV, XSLC, R, SPSS
11	Figshare	Academicgroup	https://figshare.com/	Open	open	(more than 40)	CSV, XSLX
12	SND	Swedish national data service	https://snd.gu.se/sv	permission required1	not explicit	0	-
13	eResearch South Australia	Joint venture of universities	https://data.sa.edu.au/	open	open	1	Text

Źródło: [90]

3.7. Analiza przypadków

Bazując na danych otwartych udostępnianych przez administrację i inne podmioty publiczne zbudowano już kilka usług, które pojawiły się na rynku. Przykłady zbudowanych aplikacji, serwisów i usług można wraz z opisem przeglądać m.in.:

- **Dla Polski** na stronie dane gov.pl: https://dane.gov.pl/application. Występują tutaj krótkie opisy wdrożonych usług, w dużej części opisane w powyższym raporcie. Nie ma niestety informacji o konkretnych usługach, na których zbudowane są poszczególne aplikacje i serwisy[114].
- **Dla UE** listę wdrożeń i zastosowań dla instytucji publikujących dane i wykorzystujących dane już opublikowane można przeglądać korzystając z https://www.europeandataportal.eu/en/using-data/usecases i https://data.europa.eu/euodp/apps lub opierając się na tworzonej przez społeczność liście https://opendataimpactmap.org/usecases.
- **Dla Stanów Zjednoczonych** skompletowano listę 500 firm, które w swoim modelu działania wykorzystują open data: http://www.opendata500.com/us/list/ oraz listę aplikacji na oficjalnej stronie rządowej https://www.data.gov/applications. Większość wykorzystywanych przez aplikacje zbiorów jest jednak opisana dość ogólnie.

Szczegółowiej opisując wdrożenia aplikacji można podać kilka przykładów wskazując również **aplikacje, które powstały na polskim rynku** i mogą służyć za przykłady możliwości wykorzystania udostępnionych zbiorów otwartych danych. Aplikacje, w zależności od sposobu wykorzystania open data można podzielić na 3 kategorie serwisów:

1. Wyświetlające i agregujące dane w oparciu jedynie o dane otwarte (serwis „Polska w liczbach", rys. **3.6**; Flood Alerts rys. **3.39**).
2. Wyświetlające i agregujące i przetwarzające dane z wielu źródeł, nie tylko otwartych, ale i własnych lub komercyjnych (Kanarek, rys. **3.42**).
3. Budujące własne platformy i zbiory danych, które następnie są udostępniane jako open data (Airly, rys. **3.13**).

Poniżej opisane zostaną poszczególne przykłady gotowych usług opartych na otwartych danych.

3.7.1. Fivethirtyeight (USA)

Jeżeli mowa o trzeciej kategorii serwisów, czyli tworzeniu zbiorów danych, to można do niej zaliczyć również udostępnianie już wstępnie przetworzonych otwartych zbiorów. Przykładem może być serwis dziennikarski serwis "fivethirtyeight"[115], który publikuje artykuły i analizy na

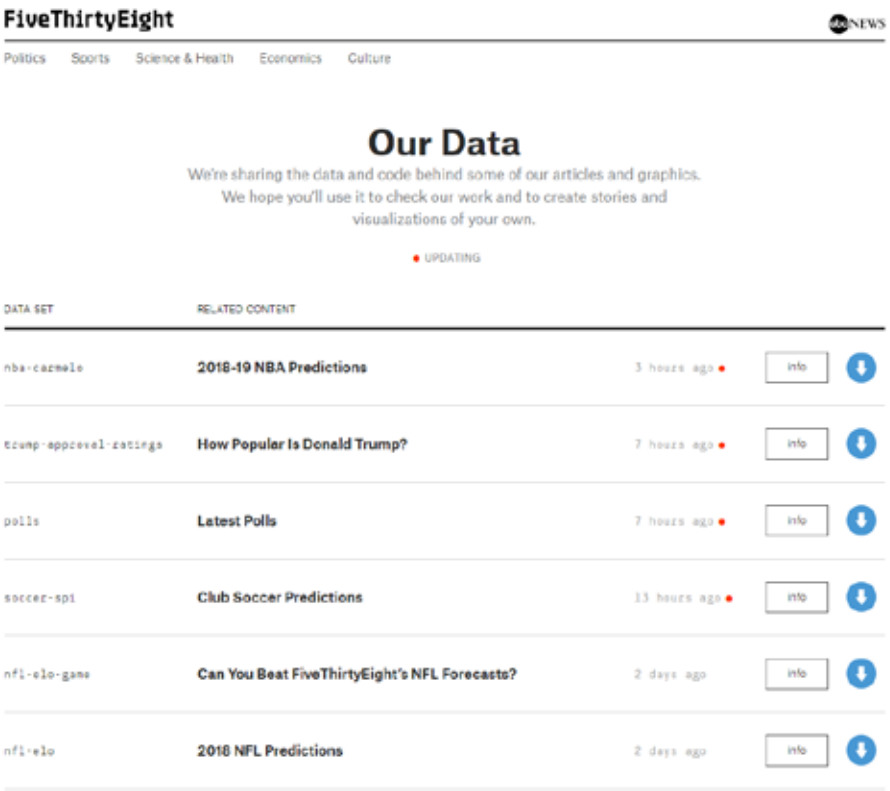
114 Wdrożenia mogą być opisane także na poziomie lokalnym – patrz Poznań http://www.poznan.pl/mim/api/wdrozenia,p,25877,38304.html.

115 https://fivethirtyeight.com/

pograniczu badań socjologicznych i statystycznych[116] z wykorzystaniem analizy danych i innowacyjnych narzędzi analitycznych. Poza artykułami opisującymi kwestie dotyczące ekonomii, polityki czy sportu, serwis korzysta także z naukowych analiz – do których wykorzystuje zwykle otwarte zbiory danych. Dane przetworzone i wykorzystane do ich przeprowadzenia serwis nieodpłatnie udostępnia na swojej stronie internetowej (rys. **3.37**). Nie jest to typowe podejście znane z open data. Mimo iż serwis nie publikuje zbiorów danych w innowacyjny sposób w perspektywie technicznej, jest to jednak innowacja na poziomie działalności przedsiębiorstwa. Udostępnione zbiory wykorzystane są jako wzmocnienie rzetelności artykułu i serwisu oraz podkreślenie transparentności wnioskowania – co stanowczo wzmacnia merytoryczną wartość serwisu publicystycznego.

Usługa oparta na: danych rządowych open data, danych pobranych z API serwisów społecznościowych i innych dostępnych źródeł.

RYSUNEK 3.37. **ZBIORY DANYCH UDOSTĘPNIONE DO ANALIZ PRZEZ SERWIS FIVETHIRTYEIGHT**



Źródło: https://data.fivethirtyeight.com/

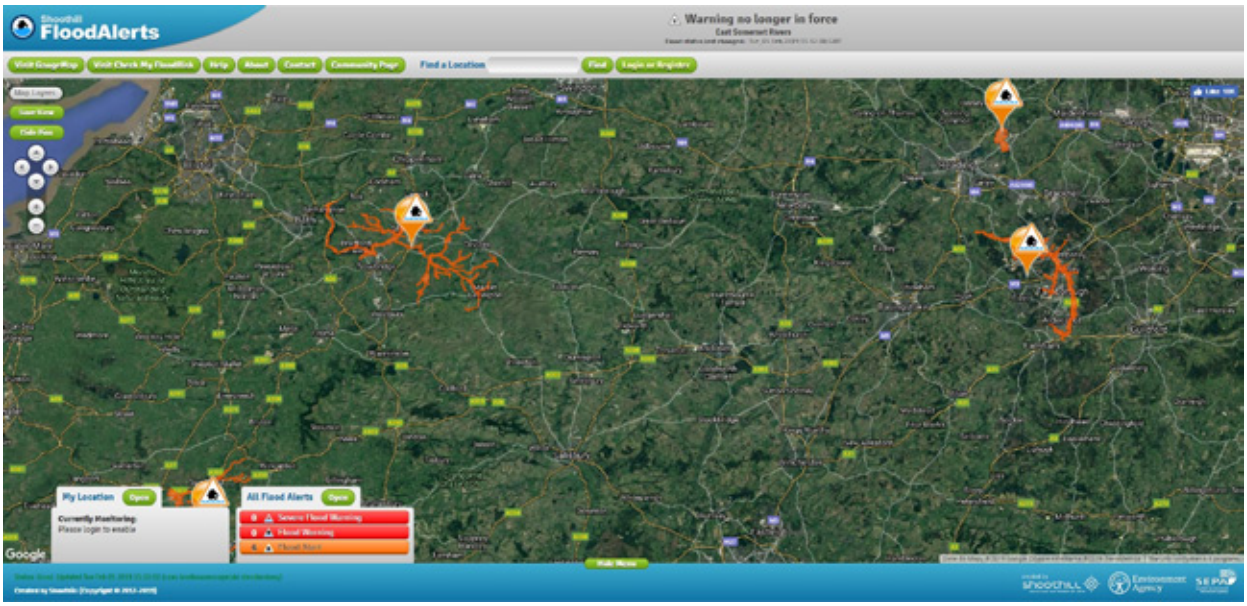
116 https://en.wikipedia.org/wiki/FiveThirtyEight

3.7.2. Flood Alerts Shoothill (UK)

Aplikacja "Flood Alerts"[117] (opis na rys. **3.39**) pochodzi z Wielkiej Brytanii. Ma ona na celu tworzenie ostrzeżeń powodziowych w oparciu o udostępniane sensory i usługi typu WMS. Serwis udostępniający mapę widoczny jest na rysunku **3.38**. Serwis wykorzystuje dane geograficzne o poziomie wód w rzekach udostępnianych przez rząd jako open data. Dane są publikowane przez wydziały zajmujące się ochroną środowiska, dane są zawsze aktualne ponieważ opierają się na odczytach z sensorów z dostępem do sieci (podobnie jak Polskie dane IMGW).

Usługa oparta na: endpoincie API udostępniającym dane o poziomie wód wydawane przez „Environment Agency" (Agencję ds. środowiska)[118]. Uwzględnione regularne odpytywanie serwisu udostępniającego poziom wód[119], co wraz z ustalonym dla danej rzeki poziomem ostrzegawczym i skutkującym powodziami pozwala na działanie aplikacji.

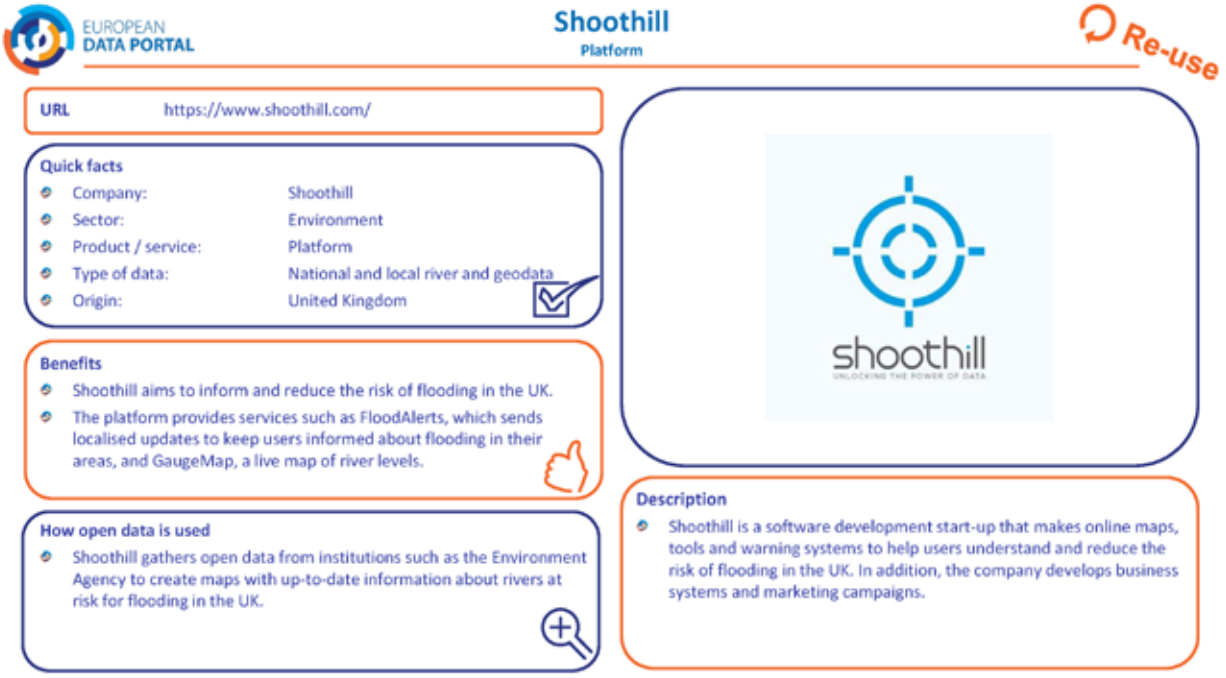
RYSUNEK 3.38. **ZRZUT EKRANU Z APLIKACJI "FLOOD ALERTS" SHOOTHILL**



Źródło: https://www.floodalerts.com/

117 https://www.floodalerts.com/
118 https: //environment.data.gov.uk/flood-monitoring/doc/reference
119 https://environment.data.gov.uk/flood-monitoring/data/readings?latest

RYSUNEK 3.39. **OPIS APLIKACJI "SHOOTHILL" NA EUROPEJSKIM PORTALU DANYCH**



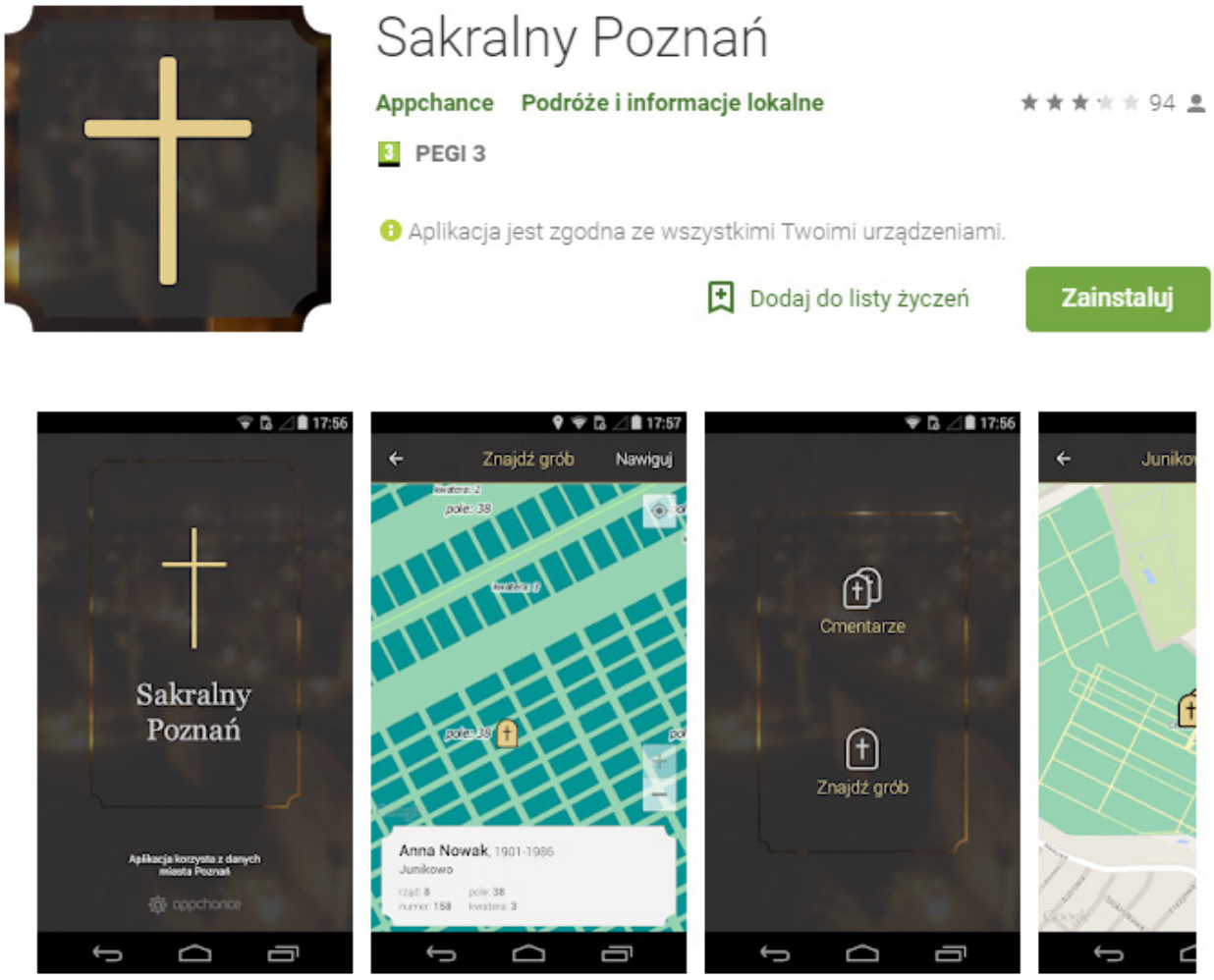
Źródło: https://www.europeandataportal.eu/sites/default/files/united_ kingdom_-_shoothill.pdf

3.7.3. Sakralny Poznań (PL)

Jako ciekawy przykład wdrożenia wskazać można aplikację "Sakralny Poznań"[120], która została oparta na otwartych danych udostępnionych przez miasto Poznań. Udostępnione za pomocą API przez poznański serwis dane, wraz z usługami opartymi na mapach pozwoliły na zbudowanie aplikacji pozwalającej na zrealizowanie bardzo prostego celu: odnalezienia grobu osób bliskich. Wyszukiwać pochowanych można poprzez imię i nazwisko, datę śmierci lub pochówku. Aplikacja, którą widać na rys. **3.40** pokazuje dostępne w aplikacji ekrany.

120 https://play.google.com/store/apps/details?id=com.appchance.graves

RYSUNEK 3.40. **APLIKACJA SAKRALNY POZNAŃ – GOOGLE PLAY STORE**

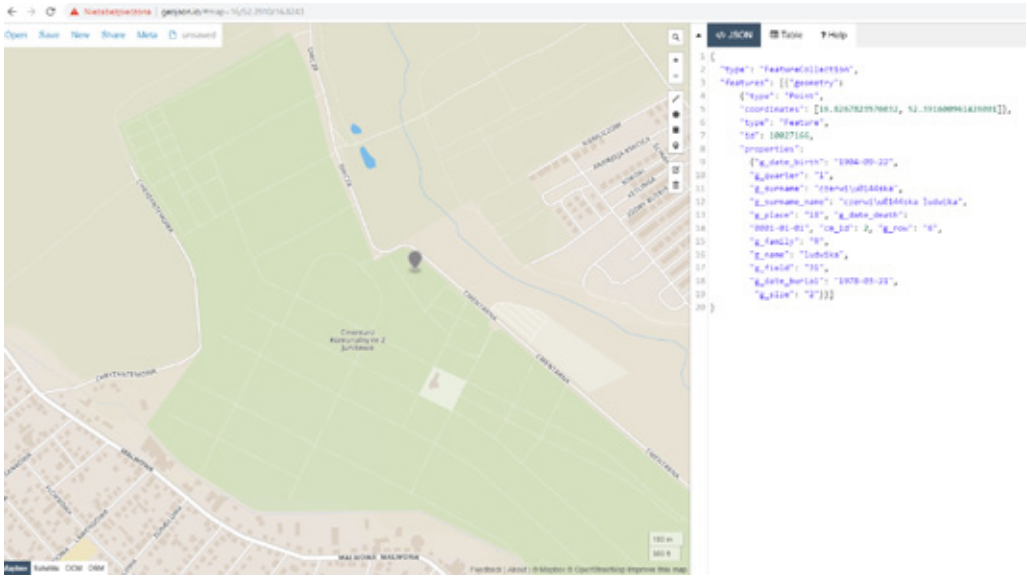


Aplikacja Sakralny Poznań powstała z myślą o osobach planujących odwiedzić groby swoich bliskich. Ma ona na celu ułatwienie odnalezienia miejsca pochówku konkretnych osob oraz wyszukanie cmentarzy i kościołów w Poznaniu.

Źródło: https://play.google.com/store/apps/details?id=com.appchance.graves

Usługa oparta na: endpoincie API udostępniany przez miasto Poznań[121]. Dane udostępniane przez serwis są w formacie GeoJson jako punkty posiadające dodatkowo dane informujące o poszczególnym miejscu pochówku. Przykład zwróconego przez API wyniku można zobaczyć w listingu **3.11**. Zwrócony wynik jest łatwy do wizualizacji, ponieważ GeoJson jest powszechnie przyjętym standardem opisu obiektów geograficznych. Przykład wizualizacji danych o obiekcie w tym formacie przez serwis geojson.io widoczny jest na rysunku **3.41**.

RYSUNEK 3.41. **PRZYKŁAD WIZUALIZACJI DANYCH O LOKALIZACJI GROBU – POZNAŃ**



Źródło: http://geojson.io/ wraz z danymi z kodu 3.11.

121 https://egov.psnc.pl/node/29#wyszukiwarka_grobow

KOD ŹRÓDŁOWY 3.11. **PRZYKŁAD WYNIKU ZWRÓCONEGO ZA POMOCĄ WYSZUKIWARKI GROBÓW UDOSTĘPNIONEJ JAKO API SERWISU HTTP://EGOV.PSNC.PL/NODE/29**


```
{"crs":
    {"type": "none",
        "properties":
            {"info": "No CRS information has been provided with this data."},
            "type": "FeatureCollection",
            "features":[
                {"geometry":
                    {"type": "Point", "coordinates": [16.8267823576032, 52.391600961426001]},
                    "type": "Feature",
                    "id": 10027166,
                    "properties":
                        {"g_date_birth": "1904-09-22",
                            "g_quarter": "1",
                            "g_surname": "czerwi\u0144ska",
                            "g_surname_name": "czerwi\u0144ska ludwika",
                            "g_place": "18",
                            "g_date_death": "0001-01-01",
                            "cm_id": 2, "g_row": "6",
                            "g_family": "R",
                            "g_name": "ludwika",
                            "g_field": "31",
                            "g_date_burial": "1978-03-21",
                            "g_size": "2"
                        }
                }]
        }
}
```


3.7.4. Kanarek (PL)

Drugą z aplikacji, którą należy wskazać jako dobre wykorzystanie wielu zbiorów danych to "Kanarek – ostrzeżenia o smogu"[122]. Aplikacja ta pozwala na wizualizację poziomu zanieczyszczenia powietrza w wielu stacjach – jest niejako nakładką graficzną na API udostępniające sensory zanieczyszczenia powietrza. Aplikacja cieszy się dużą popularnością i agreguje dane (zgodnie z opisem) ze 150 oficjalnych stacji GIOŚ, 550 LookO2, a także 350 Syngeos, 150 Luftdaten, 70 SmogTok i 35 perfect-Air. Korzysta więc nie tylko z czujników udostępnionych przez administrację, ale także prywatnych dostawców. Taki model biznesowy może wiązać się oczywiście z zagrożeniem zamknięcia dostępu do danych przez nich udostępnianych – dlatego jak w każdym przypadku, powinno brać się pod uwagę monetyzację rozwiązania w przypadku, gdyby dostęp do innych zbiorów stał się płatny.

Usługa oparta na: API GIOŚ i IMGW, opisane w raporcie (sekcja **3.2.1**) oraz API serwisu Airly (rys. **3.13**), Looko2 i podobnych dostawców.

122 https://play.google.com/store/apps/details?id= pl.tajchert.canary&hl=pl

RYSUNEK 3.42. **APLIKACJA KANAREK – GOOGLE PLAY STORE**

Źródło: https://play.google.com/store/apps/details?id= pl.tajchert.canary&hl=pl

3.7.5. Integracja wielu źródeł – ocena atrakcyjności nieruchomości (PL)

W przypadku budowy usług, które korzystają z wielu źródeł otwartych danych, ciągle istnieją na polskim rynku możliwości integracji dostępnych zbiorów i budowy usług w oparciu o nie. Jako przykład i rekomendacja sposobu myślenia przedstawiony zostanie krótko pomysł budowy aplikacji w oparciu o integracje danych z wielu źródeł. Pomysł ten pozwoli przybliżyć sposób budowy usług opartych na otwartych danych. Zadaniem zaprojektowanej aplikacji byłaby **charakterystyka obszaru geograficznego na poziomie działki**, mogąca służyć wycenie nieruchomości. Dzięki zebraniu statystyk w serwisach opartych na mapach możliwe byłoby zbudowanie serwisu pomagającego w ocenie nieruchomości.

- **Cel:** Serwis/aplikacja umożliwia ocenę atrakcyjności inwestycyjnej nieruchomości pod względem różnych statystyk.
- **Sposób działania:** Użytkownik podaje adres lub współrzędne geograficzne nieruchomości.

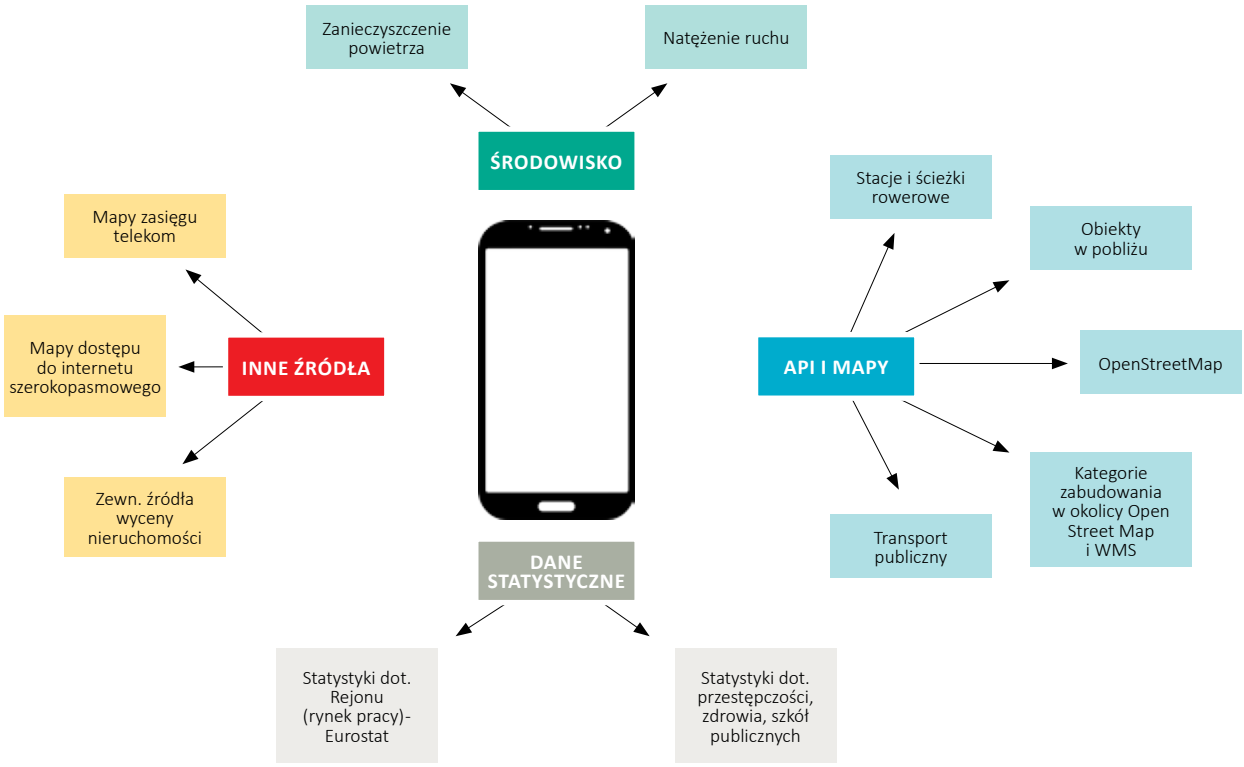
Atrakcyjność nieruchomości oparta jest na modelu przedstawionym na rysunku **3.43**, obejmującym:

- **Dane dotyczące środowiska**, w tym historyczne, wymagane jest więc zapisywanie wyników uzyskanych z API cyklicznie do bazy danych. API IMGW, dane o stanie rzek pobrane ze strony Krajowego Instytutu Meteorologii[123].
- **Dane o pobliskim natężeniu ruchu** z Google Maps[124] lub lokalnych sensorów.
- **Odległość do ważnych obiektów w pobliżu**: szkół, posterunków policji, poczty, urzędu skarbowego itd, a także potencjalnie zmniejszających atrakcyjność miejsc, takich jak aktywne lotniska[125] czy fabryki. Większość danych można uzyskać z systemów WMS albo usługi OpenStreetMap.
- **Gęstość i klasa zabudowania w pobliżu** w tym np. lesistość terenu w okolicy i liczba parków. Dane takie można wyekstrahować z OpenStreetMap na podstawie klasyfikacji terenu, tzw. "landuse"[126].
- **Transportu publicznego** i czasu potrzebnego na dojazdu do najbliższego dużego miasta, centrum miasta i stolicy województwa. Rozwiązaniem mogłoby być skorzystanie odpłatnie z API jakdojadę[127] lub bezpośrednie podłączenie do serwisów odpowiednich miast.

123 http://instytutmeteo.pl/aktualne-stany-rzek-w-polsce
124 https://developers.google.com/maps/documentation/ javascript/examples/layer-traffic
125 https://wiki.openstreetmap.org/wiki/Pl:Infrastruktura_lotnicza
126 https://wiki.openstreetmap.org/wiki/Pl:U%C5%BCytkowanie_ziemi i https:// osmlanduse.org
127 https://jakdojade.pl/public/pages/api/

- **Map zasięgu telekomunikacyjnego i dostępu do internetu szerokopasmowego**[128] i serwisy mapowe dostawców usług komórkowych lub odległość do stacji BTS[129].
- **Statystyk**, np. dotyczących rynku pracy i przestępczości w okolicy. Dane dot. przestępczości są dostepne na poziomie powiatu[130], dane dot. rynku pracy mogą być pobrane z serwisu BDL **3.25** lub na podstawie danych z serwisu „Polska w liczbach" (rys. **3.6**).

RYSUNEK 3.43. **PROPOZYCJA APLIKACJI**



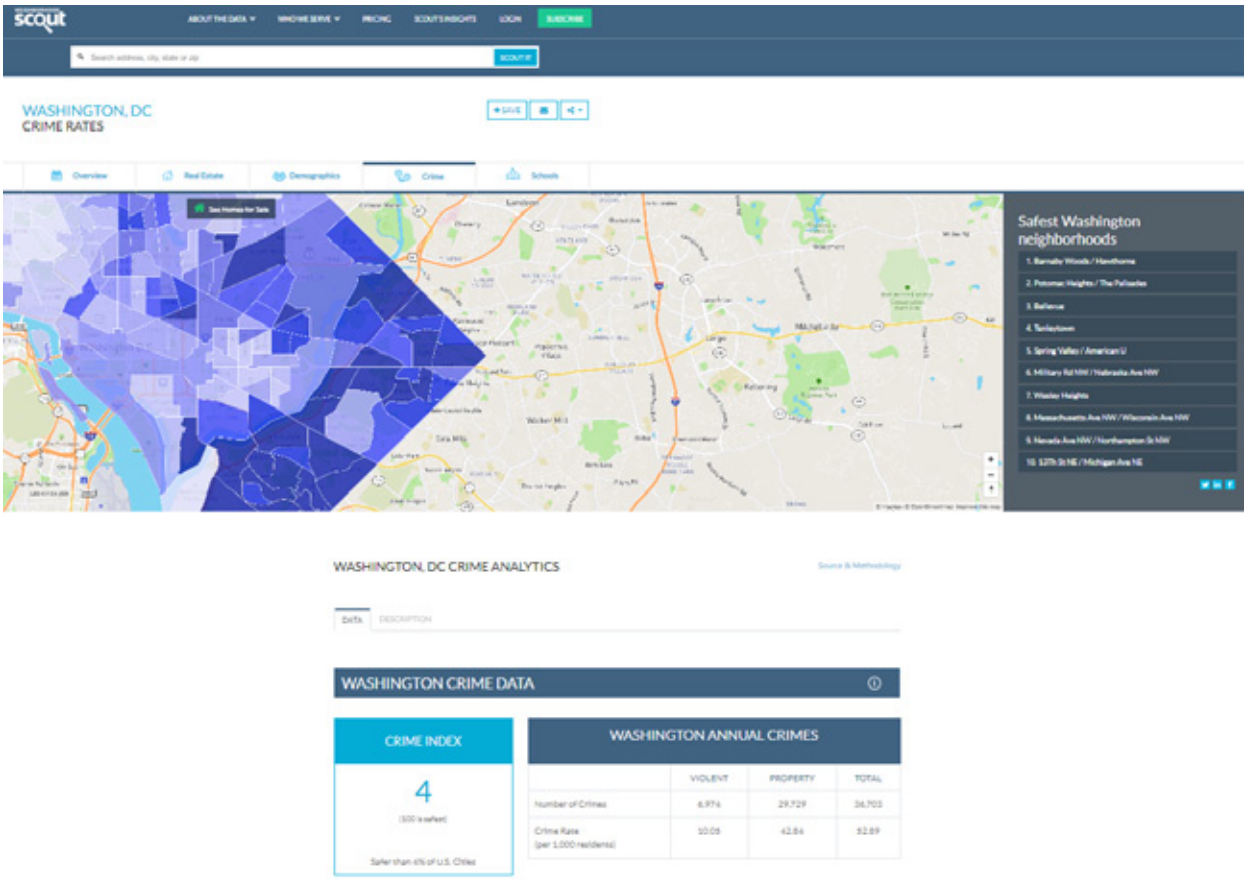
Źródło: Opracowanie własne

Usługi o podobnej tematyce można zauważyć m.in. w Stanach Zjednoczonych. Przykład to usługa "Neighborshoodscout" (rys. **3.44**), która jest w stanie podać statystyki przestępczości i bezpieczeństwa dla poszczególnych dzielnic miast. Przykłady wykorzystania geograficznych wartstw open data można także zauważyć w literaturze [**105**]. Opisany tutaj scenariusz moż-

128 https://sirs.itl.waw.pl/maps/closing_report
129 http://beta.btsearch.pl/
130 https://bdl.stat.gov.pl/BDL/metadane/cechy/szukaj?slowo= przest%C4%99pczo%C5%9B%C4%87

liwej integracji danych jest opisem bardzo ogólnym i obejmuje szeroki zakres dostępnych źródeł. Ten podany poglądowo przykład miał ona na celu wskazanie możliwie największej możliwej liczby usług w celu stworzenia funkcjonalnej aplikacji, pozwalającej na sprzedaż wyniku opracowania (atrakcyjności inwestycyjnej nieruchomości), również w systemie API na poziomie pojedynczych zapytań.

RYSUNEK 3.44. **INTERAKTYWNY SERWIS AMERYKAŃSKIEJ APLIKACJI NEIGHBORSHOODSCOUT**



Źródło: https://www.neighborhoodscout.com/dc/washington/crime

3.8. Podsumowanie

Podsumowując stan przedstawionych w powyższym rozdziale usług udostępniających dane, na podstawie opisanych usług i zasobów wskazać można krótkie rekomendacje dla przedstawionych treści. W treści tego rozdziału:

1. Wskazane zostały dane udostępniane przez polską administrację publiczną, wraz ze wskazaniem kategorii interesujących dla przedsiębiorców i realnych przykładów użycia tych zasobów.
2. Zaprezentowano europejskie i ogólnoświatowe inicjatywy open data, wraz z wnioskami i klasyfikacją Polski we wszystkich z nich. Wskazano możliwe scenariusze wykorzystania samych rankingów i przykładów w nich opisanych, które mogą być przydatne dla przedsiębiorców oraz zasoby danych (jak dane Eurostat) możliwe do wykorzystania.
3. Wskazano źródła danych otwartych ze źródeł niepublicznych. Scharakteryzowano dostawców i inicjatywy oraz najważniejsze zbiory charakteryzujące się potencjałem możliwym do wykorzystania. Większość opisanych jednak przykładów opiera się na przetwarzaniu danych publicznych już dostępnych w innych źródłach.
4. Opisując powyższe dane i usługi zwrócono uwagę na rzetelność i aktualność udostępnionych zbiorów, wskazano, czy jakość danych nie stanowi przeszkody do ich wykorzystania w przypadku poszczególnych zasobów.
5. Zaproponowano obszary i zasoby szczególnie interesujące dla przedsiębiorców, skupiając się na tych obszarach, które pozwalają na kreowanie innowacyjnych usług w skali światowej.

Przechodząc do **rekomendacji dotyczących wykorzystania opisanych zasobów danych przez przedsiębiorców** jako najbardziej wartościowe wskazujemy **usługi WMS** i **zbiory statystyczne (API BDL, Teryt i dane Eurostat)** udostępniane przez GUS i inne instytucje pozwalają na tworzenie wielu analiz o charakterze geograficznym, które mogą służyć zbudowaniu innowacyjnych usług. Przykład infrastruktury WMS dla danych geograficznych także świetnie przedstawia możliwości budowania i udostępniania platformy i standardów. Samorządy lokalne mogą wykorzystywać działającą usługę na poziomie państwowym do umieszczania w niej danych o charakterze lokalnym (np. plany zagospodarowania przestrzennego). Wykorzystanie platform o podobnym stylu budowy, gdzie architektura, format (a czasami i serwer/wdrożenie) następuje po stronie administracji krajowej, może być najlepszym sposobem wdrażania usług open data na poziomie lokalnym. W perspektywie oceny samych zasobów także usługa **dane.gov.pl** (podrozdział **3.2.1**) jest inicjatywą idącą w bardzo dobrym kierunku – w szczególności, że pozwala na udostępnianie katalogów danych, a nie tylko pojedynczych zbiorów.

Przykładowo, wykonanie badania rynku w zakresie: liczby przedsiębiorstw o podobnym charakterze, poziomu bezrobocia i umiejscowienia geograficznego centrów logistycznych czy

sprzedażowych może stać się możliwe w oparciu o wyżej wymienione usługi. Wykorzystanie dodatkowo usług otwartych o charakterze społecznościowym (jak OpenStreetMap, sekcja **4.3.7**), pozwala na zbudowanie analiz skrojonych pod konkretnego przedsiębiorcę. Przedstawione w tym rozdziale analizy przypadków i przykłady wykorzystania usług mogą pomóc nakierować i zachęcić twórców przyszłych innowacyjnych usług w zakresie open data do skorzystania z dobrze wyeksponowanych i dostępnych zbiorów wysokiej jakości.

W kontekście istniejących usług należy skupić się na **zwiększeniu zakresu dostępnych danych dynamicznych** (udostępnionych jako API) o niskim poziomie agregacji. **Bardzo ważnym elementem jest także poszerzenie katalogów danych zależnych od instytucji publicznych** (w rozumieniu planów zagospodarowania, danych z sektora szkolnictwa, zatrudnienia, statystyk policyjnych). W tej chwili brakuje w katalogach danych GUS i dane.gov.pl odpowiednich inicjatyw i programów do otwierania danych publicznych w innych obszarach, które powinny być dostępne dla obywateli (np. statystyki dotyczące przestępczości w mniej zagregowanej formie), lub pozwoliłyby na budowanie innowacyjnych usług (np. aktualne dane dotyczące zagospodarowania przestrzennego i specyfiki budynków znajdujących się na danym obszarze geograficznym). Obecne w tej chwili dane GUS i dane.gov.pl mają dobrą podstawę techniczną, ale nie do końca pozwalają na budowanie innowacyjnych usług.

Wiele zasobów danych dostępnych w różnych instytucjach w Polsce korzysta z instytucji "głębokiego ukrycia", np. dane z ksiąg wieczystych czy KRS. Oznacza to, że są one dostępne publicznie, jednakże zwykle brak jest możliwości ich otrzymania w cyfrowej formie, co bardzo mocno blokuje potencjał rozwoju w zakresie budowy innowacyjnych usług bazujących na tych danych. W zakresie danych publicznych strona rządowa powinna jasno określić, jakie zbiory danych są wyłączone z użytku, bądź wymagają szczególnych uprawnień, zamiast ograniczać dostęp poprzez nieproporcjonalne zwiększanie kosztów przetwarzania dla przedsiębiorców. Powinna być też ustalona jasna polityka korzystania z danych mechanicznych i pochodzących z sensorów.

Rekomendacje dotyczące tematyki budowanych usług stworzone przez autorów tego raportu są spójne ze wskazaniami aktualizacji dyrektywy PSI (podrozdział **3.3.2**) i najnowszymi trendami badań w obszarze smart cities (sekcje **3.2.3** i **3.5.3**). Zainteresowanie inteligentnymi sensorami i działaniami w zakresie optymalizacji zużycia energii, działaniami proekologicznymi w oparciu o dane, transportem i zagospodarowaniem przestrzennym z wykorzystaniem innowacyjnych źródeł danych, może przełożyć się na duży wzrost innowacyjności projektowanych usług.

4 Crowdsourcing – społeczność dla społeczności

4.1. Wprowadzenie

Model tworzenia treści przez rozproszoną społeczność szybko zmaterializował się jako model biznesowy. Firmy często zlecają internautom realizację określonego zadania, które stanowi zbyt duże wyzwanie dla pojedynczego zespołu czy nawet firmy. Poprzez analogię do outsourcingu, zjawisko to zostało określone jako crowdsourcing. Dziś jest to globalnie rozpoznawalny fenomen, który stał się podstawą działania wielu firm. Na przykład Kaggle jest platformą do organizacji konkursów z analityki danych, a działająca non profit Khan Academy ma misję dostarczania za darmo materiałów edukacyjnych najwyższej jakości, OpenStreetMap to darmowa mapa świata tworzona przez internautów, Waze jest aplikacją mobilną wykorzystująca crowdsourcing do zbierania informacji o ruchu i zdarzeniach drogowych.

Niektóre bazy wiedzy (np. DBpedia, Web Data Commons) udostępniają swoje dane w formacie RDF (Resource Description Framework). Każdy rekord w takim formacie reprezentowany w postaci tzw. semantycznej trójki, która przypomina zdanie złożone z podmiotu, orzeczenia i dopełnienia [**172**]. Trójki te można zapisywać przy użyciu znaczników XML. W ramach RDF dokument zawiera stwierdzenia typu: określony podmiot (np. rzecz, osoba, dokument itp.) pozostaje w pewnej relacji (np. odpowiednio „kosztuje", „mieszka w", „napisany przez") z innym podmiotem (rzeczą, osobą, dokumentem itp.). Dzięki temu w naturalny sposób da się opisać większość danych przetwarzanych przez komputery. „Podmiot" oraz „dopełnienie" są wyrażone przez URI (Universal Resource Identifier – uniwersalne oznaczenie zasobu), czyli globalnie unikatowe identyfikatory.

Przy pomocy specjalnych narzędzi można dowiedzieć się, czy strona internetowa zawiera dane semantyczne w formacie RDF. Np. narzędzie od W3C[1] pozwala otrzymać wszystkie trójki w formacie RDF/XML, JSON-LD, Turtle czy zwykłego tekstu po wprowadzeniu adresu strony WWW. Na rys. **4.1** pokazany przykład ekstrakcji semantycznych trójek przy pomocy tego narzędzia dla strony o filmie „Avatar" z serwisu Filmweb.

1 https://www.w3.org/2012/sde/

RYSUNEK 4.1. **PRZYKŁAD EKSTRAKCJI DANYCH W FORMACIE RDF ZE STRONY OPISUJĄCEJ FILM „AVATAR" Z SERWISU FILMWEB**

<.../Avatar>	<http://rdf.data-vocabulary.org/#starring>	<.../person/Sigourney.Weaver> .
<.../Avatar>	<http://rdf.data-vocabulary.org/#starring>	<.../person/CCH+Pounder-9197> .
<.../Avatar>	<http://rdf.data-vocabulary.org/#starring>	<.../person/Joel+David+Moore-144362> .
<.../Avatar>	<http://rdf.data-vocabulary.org/#starring>	<.../person/Giovanni+Ribisi-4353> .
<.../Avatar>	<http://ogp.me/ns#type>	video.movie .
<.../Avatar>	<http://rdf.data-vocabulary.org/#starring>	<.../person/Stephen+Lang-4679> .
<.../Avatar>	<http://rdf.data-vocabulary.org/#starring>	<.../person/Michelle+Rodriguez-40280> .
<.../Avatar>	<http://ogp.me/ns#site_name>	Filmweb .
<.../Avatar>	<http://ogp.me/ns#image>	https://ssl-gfx.filmweb.pl/po/91/13/299113/7332755.3.jpg .
<.../Avatar>	<http://ogp.me/ns#url>	.../Avatar .
<.../Avatar>	<http://rdf.data-vocabulary.org/#starring>	<.../person/Sam+Worthington-53674> .
<.../Avatar>	<http://ogp.me/ns#title>	Avatar .
<.../Avatar>	<http://rdf.data-vocabulary.org/#starring>	<.../person/Zoe+Saldana-40559> .

Źródło: https://www.filmweb.pl/Avatar. W celu lepszej widoczności danych „https://www.filmweb.pl" zastąpiono „...".

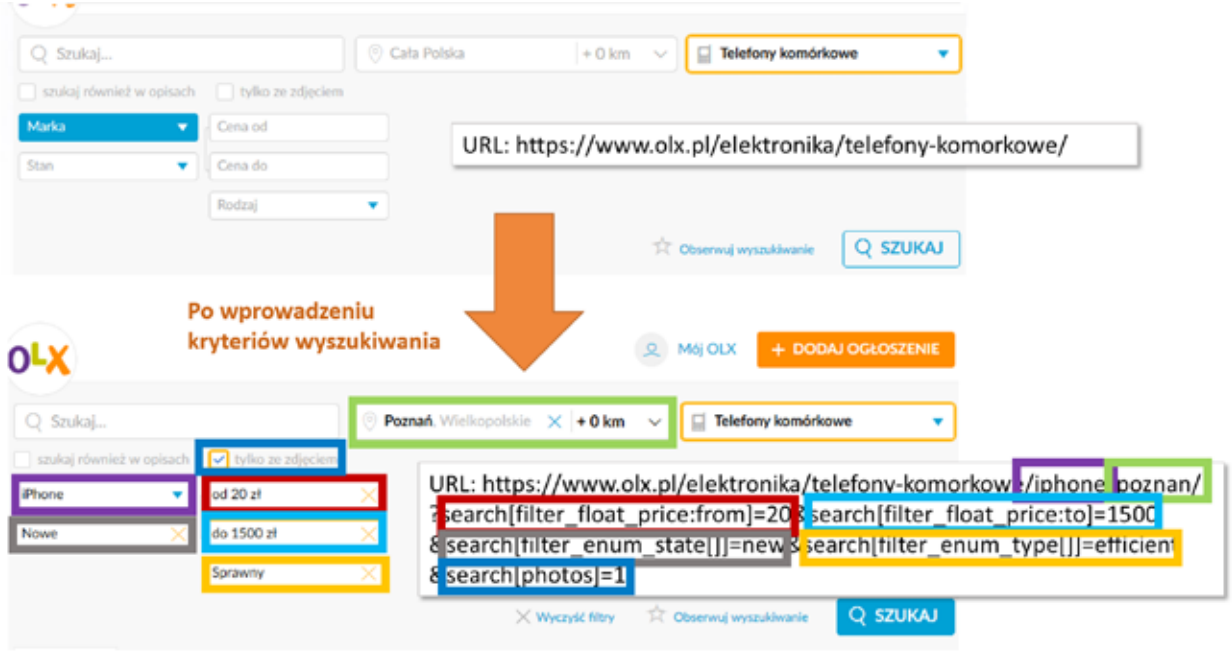
4.2. Web scrapping (Deep Web)

Wraz z popularnością internetu rośnie ilość treści przechowywanych w wewnętrznych bazach danych, do których wyszukiwarki nie zawsze docierają – mówimy, że Internet został „pogłębiony" dzięki ogromnym bazom danych online [**68**]. Treść ukryta za formularzami HTML od dawna uznawana jest za znaczną lukę w zasięgu algorytmów indeksujących globalną sieć [**112**]. Te głębokie treści internetowe mogą mieć znacznie lepszą jakość niż te w sieci WWW [**84**].

W celu uzyskania danych z Deep Web zazwyczaj należy wykonywać zapytania bezpośrednio w przeglądarce. Wiele aplikacji związanych z ekstrakcją treści z głębokiego internetu opiera się na zrozumieniu schematu tych interfejsów zapytań. Schemat musi obejmować mapowania elementów wejściowych i etykiet, typów danych, poprawnych wartości oraz zakresów wartości danych wejściowych. Dodatkowo, aby wyekstrahować te ukryte dane, schemat musi zawierać wiele informacji związanych z przesyłaniem formularzy, takich jak pliki cookie (tzw „ciasteczka") oraz typy działań do zapytania [**84**].

Z technicznego punktu widzenia, interfejsy zapytań są częścią strony HTML zawartej pomiędzy tagami <form> oraz </form>. Interfejs formy może być umieszczony na specjalnej stronie i jego struktura może się różnić w zależności od obszaru wyszukiwanych danych. Na przykład, w przypadku telefonów komórkowych możemy mieć formularz wyszukiwania inny niż dla samochodów osobowych, ponieważ dla każdego z tych typów towaru mogą występować różne charakterystyki. Na rys. **4.2** pokazany został interfejs wyszukiwania dla telefonów komórkowych na portalu OLX. Warto zwrócić uwagę na to, że po wprowadzeniu niektórych kryteriów wyszukiwania mamy dostęp do strony z innym adresem niż strona z pustym formularzem. Zmiana każdego z parametrów wyszukiwania może zmienić adres strony URL z wynikami.

RYSUNEK 4.2. **FORMULARZ WYSZUKIWARKI TELEFONÓW NA PORTALU OLX PRZED ORAZ PO WPROWADZENIU KRYTERIÓW**



Źródło: https://www.olx.pl/

4.2.1. Ekstrakcja informacji

W idealnym internecie dane byłyby dostarczane do klientów w wybranym przez nich formacie: CSV, XML, JSON itp. W realnym Internecie nie wszystkie dane są dostępne w postaci strukturalnej – zazwyczaj mamy do czynienia z dokumentami w formatacie HTML, przeznaczonym raczej do wyświetlania danych niż ich wymiany.

Transfer danych pomiędzy programami może się odbywać za pośrednictwem struktur dobrze uporządkowanych, łatwych do odczytu. Z drugiej strony, maszynowe struktury danych są bardzo często nieczytelne dla człowieka, który wymaga odpowiedniego formatowania, komentarzy oraz innych dodatkowych informacji.

W sytuacji, kiedy mamy na wyjściu dane, które są dostępne jedynie w przyjaznym człowiekowi formacie, technika screen scraping (dosł. „zeskrobywanie z ekranu") staje się jedynym zauto-

matyzowanym sposobem pozyskania danych. Innymi słowy, screen scraping to technika, za pomocą której program komputerowy wydobywa dane z wyjścia innego programu. Ta technika zawiera także skomputeryzowane przetwarzanie kodu HTML na stronach internetowych.

Rozpatrzmy prosty przykład strony, z której mamy zamiar wyekstrahować informacje (patrz kod **4.1**).

KOD ŹRÓDŁOWY 4.1. **PRZYKŁAD STRONY HTML DO EKSTRAKCJI**

```
1   <!DOCTYPE html>
2   <html>
3       <head>
4           <title>Nazwa dokumentu</title>
5       </head>
6       <body>
7           <h1>Nagłówek 1-go poziomu</h1>
8           <p>Tekst...</p>
9       <body>
10  </html>
```

Każdy <tag> obsługuje blok wewnątrz strony:

- Dokumenty HTML muszą rozpoczynać się od deklaracji typu <!DOCTYPE html>.
- Dokument HTML znajduje się pomiędzy <html> i </html>.
- Deklaracja meta danych oraz skryptów dokumentu HTML znajduje się pomiędzy <head> i </head>. W tym obszarze jest umieszczony tag <title>, który deklaruje nazwę dokumentu dla przeglądarki oraz systemów wyszukiwawczych.
- Widoczna część dokumentu HTML znajduje się pomiędzy znacznikami <body> i </body>.
- Nagłówki tytułów definiuje się między znacznikami <h1> do <h6>.
- Akapity są definiowane za pomocą znacznika <p>.

Innymi użytecznymi tagami mogą być również:

- <a> – dla hiperłączy,
- <table> – dla tabel,
- <tr> – dla wierszy tabel, • <td> – dla kolumn tabeli.

Ponadto znaczniki HTML czasami zawierają atrybuty id lub class. Atrybut id określa unikalny identyfikator dla znacznika HTML, a wartość musi być unikalna w dokumencie HTML. Atrybut class służy do definiowania stylów równych dla tagów HTML z tą samą klasą. Użycie identyfikatorów i klas pomaga zlokalizować dane, które należy wyekstrahować.

Dobrze zaprogramowany algorytm do ekstrakcji stron HTML musi nie tylko służyć do przetwarzania interesujących danych, lecz także być w stanie odrzucić niechciane informacje oraz formatowanie. Rozpatrzymy przykłady opracowane przy użyciu języka Python, wraz z biblioteką BeautifulSoup.

Dla prostej strony z naszego przykładu chcemy wyekstrahować tekst, który jest umieszczony w akapicie – znacznik <p>. Do ekstrakcji można wtedy użyć kodu **4.2**.

KOD ŹRÓDŁOWY 4.2. **KOD ALGORYTMU DO EKSTRAKCJI TEKSTU Z PRZYKŁADU KODU 4.1**

```
1    from bs4 import BeautifulSoup
2    with open("example.html", "r") as plik:
3        kod = plik.read()
4    soup = BeautifulSoup(kod, 'html.parser')
5    tekst = soup.find('p').text
6    print(tekst)
```

Przed rozpoczęciem tworzenia algorytmu do ekstrakcji informacji z określonych serwisów internetowych, należy wziąć pod uwagę następujące zasady:

- Należy sprawdzić warunki korzystania z witryny (m.in. przeczytać oświadczenia o legalnym korzystaniu z danych). Zazwyczaj zbierane przez nas dane nie powinny być wykorzystywane do celów komercyjnych.
- Nie należy zbyt agresywnie pobierać danych ze strony internetowej za pomocą programu, ponieważ może to spowodować odmowę świadczenia usług przez aplikację. Liczba dopuszczalnych zapytań w jednostce czasu jest zazwyczaj definiowana przez źródło.
- Układ witryny może się zmieniać od czasu do czasu, zatem należy pamiętać, aby w razie konieczności ponownie przeanalizować strukturę stron witryny i poprawić kod.

4.2.2. Interesujące źródła

Popularne w Polsce zasoby głębokiego Internetu (w nawiasach zaznaczono miejsce w rankingu najczęściej odwiedzanych stron WWW w Polsce[2]):

- http://allegro.pl (5 miejsce) – Polski internetowy serwis aukcyjny.
- http://olx.pl (8 miejsce) – serwis dostarcza informacji o aktualnych ofertach kupna, sprzedaży, wymiany towarów i usług.
- http://aliexpress.com (13 miejsce) – to globalny rynek detaliczny skierowany do konsumentów na całym świecie

2 https://www.alexa.com/topsites/countries/PL

- http://otomoto.pl (17 miejsce) - ogłoszenia dotyczące samochodów osobowych, dostawczych, ciężarowych i motocyklów.
- http://filmweb.pl (19 miejsce) - największa strona o filmach w Polsce i jedna z największych baz filmowych na świecie.
- http://pracuj.pl (50 miejsce) – oferty pracy w Polsce i za granicą.

4.2.3. Dostępne frameworki

Frameworki do ekstrakcji treści stron HTML są bardziej interesujące dla firm, które nie mają kompetencji czy zasobów, aby zajmować się samodzielnym pisaniem skryptów. Dobrym przykładem jest serwis Import.io[3], który posiada różne narzędzia do:

- identyfikacji adresów URL, w którym znajdują się dane,
- ekstrakcji wyświetlanej lub ukrytej zawartości z dowolnego miejsca w Internecie,
- przygotowania wyodrębnionych danych, a także eksploracji, oceny i udoskonalenia danych,
- integracji przygotowanych danych z biblioteką interfejsów API, aby zapewnić integrację z wewnętrznymi systemami biznesowymi lub dostarczyć je do dowolnego repozytorium w celu opracowania zaawansowanych funkcji analitycznych,
- wykorzystania przygotowanych danych do budowania wykresów w celu znalezienia odpowiedzi oraz zdobycia wiedzy, a także analizowania danych za pomocą zmian, porównań oraz niestandardowych raportów.

W szczególności warto zwrócić uwagę na dedykowane narzędzie do ekstrakcji stron internetowych https://www.import.io/builder/data-extraction/. Inne przydatne narzędzia:

- https://doc.scrapy.org/en/latest/
- https://www.screen-scraper.com/

3 https://www.import.io/

4.3. Rozproszone pozyskiwanie danych

4.3.1. Serwisy wiki

Serwis wiki stanowi zbiór powiązanych ze sobą stron internetowych, w których treść można tworzyć i zmieniać z poziomu przeglądarki internetowej, a wykorzystuje się je np. do pracy nad wspólnymi projektami.

Strony typu wiki to również platformy wymiany wiedzy, promujące jej tworzenie poprzez wzajemną współpracę [**82**]. Różnica między wiki a blogiem polega na tym, że witryny wiki są zaprojektowane we współpracy z grupami użytkowników, a każdy z członków może dowolnie edytować publikowane treści. W ramach takich serwisów dostępne są również fora dyskusyjne dla każdej strony. Jedną z głównych organizacji, która zarządza różnymi serwisami typu wiki, jest Fundacja Wikimedia. Obecnie prowadzi ona 702 aktywne projekty w różnych językach [**195**]. Tabela **4.1** przedstawia 20 największych projektów fundacji Wikipedia z uwzględnieniem wersji językowych.

Serwisy wiki mogą działać nie tylko jako ogólnodostępne zbiory informacji, ale również jako korporacyjne bazy wiedzy. Takie serwisy zapewniają większą przejrzystość wszystkim procesom w organizacjach czy firmach. Dodatkowo serwisy wiki umożliwiają pracownikom współpracę w zakresie komunikacji i dzielenia się informacjami, aktywnie rozwijać nowe idee, a także umożliwiają integracje pracowników firmy oraz szybkie wyszukiwanie ekspertów z różnych dziedzin.

Istnieje wiele możliwości tworzenie własnych serwisów wiki przy pomocy ogólnodostępnego oprogramowania. Poniżej znajduje się lista z opisem niektórych z nich:

- **MediaWiki**[4] – to platforma zbudowana przy użyciu języka PHP na potrzeby tworzenia dużych projektów. Jest to oprogramowanie, na którym działa Wikipedia i inne projekty Fundacji Wikimedia[5].

4 https://www.mediawiki.org
5 https://wikimediafoundation.org/our-work/wikimediaprojects/

TABELA 4.1. **LISTA 20 NAJWIĘKSZYCH PROJEKTÓW FUNDACJI WIKIMEDIA POD WZGLĘDEM LICZBY ARTYKUŁÓW Z OKREŚLENIEM WERSJI JĘZYKOWYCH**

PROJEKT	JĘZYK	ARTYKUŁY	STRONY	EDYCJE	UŻYTKOWNICY
commons.wikimedia		48 878 991	66 380 600	317 353 404	7 205 568
en.wiktionary	angielski	5 748 814	6 359 655	50 219 716	3 456 092
en.wikipedia	angielski	5 709 436	45 759 359	852 931 885	34 376 843
ceb.wikipedia	cebuański	5 381 698	8 964 659	23 687 176	50 728
mg.wiktionary	malgaski	5 099 433	5 185 980	25 938 861	6 793
sv.wikipedia	szwedzki	3 771 367	7 690 453	43 411 792	628 559
fr.wiktionary	francuski	3 334 283	3 630 632	25 459 986	234 489
de.wikipedia	niemiecki	2 215 479	6 211 862	179 396 475	2 987 641
fr.wikipedia	francuski	2 037 063	9 695 952	151 378 202	3 197 547
nl.wikipedia	holenderski	1 940 752	3 994 645	52 019 551	939 362
ru.wikipedia	rosyjski	1 494 050	5 775 523	94 537 829	2 359 692
es.wikipedia	hiszpański	1 467 398	6 451 679	109 908 672	5 122 582
it.wikipedia	włoski	1 458 461	5 908 572	99 128 529	1 707 762
pl.wikipedia	polski	1 297 520	2 895 965	54 191 992	906 440
war.wikipedia	warajski	1 263 158	2 876 525	6 193 052	37 830
vi.wikipedia	wietnamski	1 187 698	13 617 710	42 228 218	618 196
ja.wikipedia	japoński	1 118 856	3 315 257	69 582 586	1 373 266
zh.wikipedia	chiński	1 020 594	5 507 109	50 820 337	2 574 037
pt.wikipedia	portugalski	1 004 205	4 693 072	52 855 502	2 140 446
ru.wiktionary	rosyjski	982 771	1 416 751	9 970 031	202 151

Źródło: opracowanie własne na podstawie [195].

- **Tiki**[6] – platforma do utworzenia serwisów wiki, która dotąd została pobrana ponad milion razy przez firmy, rządy, organizacje non-profit i osoby na całym świecie. Popularność platformy wynika z liczby funkcji, jakie proponuje, ponieważ umożliwia także tworzenie stron blogów, forów, kanałów RSS oraz ankiet.
- **DokuWiki**[7] – oprogramowanie wiki, którego możliwości są bardzo zbliżone do Tiki i MediaWiki, pomimo braku niektórych zaawansowanych funkcji. Najważniejszą zaletą DokuWiki jest łatwość użytkowania.

Do utworzenia własnych serwisów wiki służą specjalne serwisy, które udostępniają platformę oraz serwer do jej działania za darmo. Jednym z takich serwisów jest Wikia[8]. Ten serwis obecnie znajduje się w rankingu 50 najczęściej

6 https://tiki.org
7 https://www.dokuwiki.org
8 http://wikia.com

odwiedzanych stron internetowych na świecie [**4**]. Za pośrednictwem Wikia działa ponad 385 tys. różnych encyklopedii, w których ogólna liczba stron wynosi ponad 50 mln [**51**]. Najczęściej w ramach danego serwisu tworzone są bazy wiedzy na temat gier wideo, filmów, muzyki, komiksów. Tabela **4.2** przedstawia listę 20 największych projektów w ramach serwisu Wikia pod kątem liczby artykułów.

TABELA 4.2. **LISTA 20 NAJWIĘKSZYCH PROJEKTÓW W RAMACH SERWISU WIKIA POD WZGLĘDEM LICZBYARTYKUŁÓW**

PROJEKT	ARTYKUŁY	STRONY	EDYCJE	UŻYTKOWNICY
respuestas.wikia	2 483 083	4 486 350	5 720 974	15 441 276
colors.wikia	2 242 280	2 242 981	2 245 234	16 580 738
lyrics.wikia	2 010 752	3 224 161	31 599 990	15 530 860
answers.wikia	1 143 393	2 079 160	7 123 589	15 430 860
speedydeletion.wikia	766 115	1 050 302	1 148 994	15 430 860
lt.biologija.wikia	583 411	1 392 322	2 075 290	16 374 841
techteam-qa6.wikia	344 179	589 526	1 032 498	7 903 436
scratchpad.wikia	282 887	449 211	2 521 285	15 420 104
familypedia.wikia	246 415	619 025	1 403 716	15 530 860
military.wikia	240 593	652 830	4 160 249	15 441 276
frag.wikia	227 211	494 517	839 458	15 430 860
marvel.wikia	209 329	1 036 637	4 588 626	15 441 276
ru.vlab.wikia	204 717	314 798	476 486	16 626 422
respostas.wikia	192 653	512 139	836 557	16 638 441
eq2.wikia	178 678	287 332	897 389	15 420 104
reponses.wikia	178 134	468 287	1 113 583	15 463 788
crossgencomicsdatabase.wikia	147 165	302 818	1 507 286	8 662 262
starwars.wikia	144 303	476 583	7 799 848	15 441 276
prowrestling.wikia	110 142	437 341	1 440 067	15 420 104
yugioh.wikia	107 584	557 680	3 969 504	15 441 276

Źródło: opracowanie własne na podstawie podstawie [197].

Innym przykładem serwisu, który umożliwia stworzenie własnych encyklopedii, jest Gamepedia[9]. Serwis umożliwia tworzenie baz wiedzy na temat gier wideo. Obecnie zawiera ponad 2000 różnych encyklopedii, z ponad 5 mln artykułami edytowanymi przez ponad 1,2 mln użytkowników.

9 https://www.gamepedia.com/

Niektóre serwisy wiki działają na własnych platformach. Na przykład Baidu Bake[10] jest chińską encyklopedią posiadającą ponad 15 mln stron, które były dotąd edytowane ponad 144 mln razy. Całkowita liczba zarejestrowanych użytkowników w tym serwisie to ponad 6,5 mln.

Inne przykłady serwisów wiki:

- OmegaWiki, duży, współredagowany słownik wielojęzyczny.
- Wikisłownik, projekt do tworzenia wielojęzycznego słownika.
- Wikicytaty, bezpłatne kompendium online z pozyskanych cytatów od znanych osób i twórczych dzieł w różnych językach.

4.3.2. Wikipedia

Jednym z najbardziej znanych przykładów źródła współtworzonego przez wiele osób jest Wikipedia. Zgodnie z jej zasadami informacja może być dostarczana przez każdego, również przez anonimowych użytkowników. Wikipedia jest popularnym przykładem serwisów wiki i często uważa się ją za projekt crowdsourcingowy [**28**].

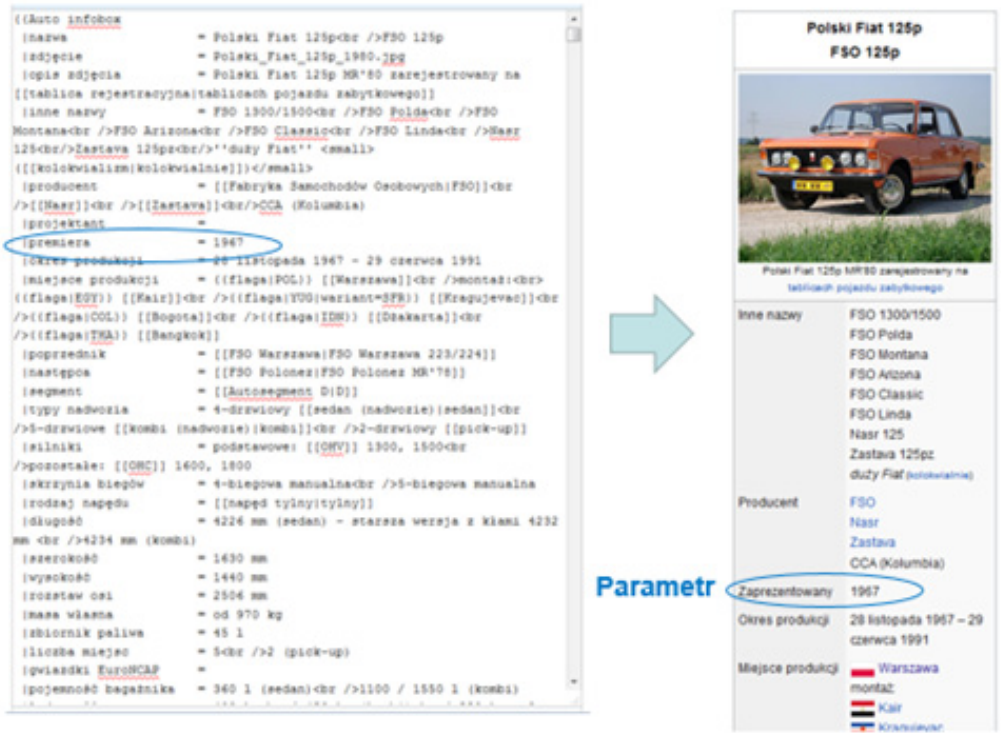
Od czasu powstania Wikipedia zdobyła pozycję jednego z ważniejszych źródeł ogólnodostępnej informacji encyklopedycznej. Jej cechą charakterystyczną jest to, że jest ona współtworzona przez wielu użytkowników. Obecnie Wikipedia jest na piątym miejscu w rankingu najczęściej odwiedzanych stron w Internecie, ustępując tylko Google, YouTube, Facebook oraz Baidu.

Koncepcja Wikipedii jest dość prosta: otwarta encyklopedia, którą może edytować każdy. Została ona uruchomiona 15 stycznia 2001 roku. Wikipedia jest stworzona przede wszystkim dla ludzi, które chcą lepiej poznać swoją historię, społeczeństwo oraz kulturę. Zarówno instytucje badawcze, jak i firmy mogą nieodpłatnie korzystać z tej encyklopedii do poszerzenia wiedzy oraz polepszenia technologii.

Obecnie Wikipedia zawiera ponad 48 mln artykułów w około 300 różnych językach [**196**]. Największa jest angielska (EN) wersja językowa, która zawiera ponad 5,5 mln artykułów. Do jednych z najbardziej rozwiniętych wersji językowych należą również niemiecka (DE) z ponad 2 mln artykułami, a także francuska (FR), rosyjska (RU), polska (PL) z ponad 1 mln artykułów każda. W Wikipedii, obok informacji na temat znanych osób, miast czy wydarzeń, znajdują się treści związane z produktami, takimi jak filmy, samochody, telefony komórkowe. Każdy produkt może być opisany w różnych językach.

10 https://baike.baidu.com/

RYSUNEK 4.3. **INFOBOKS OPISUJĄCY SAMOCHÓD (Z LEWEJ STRONY – KOD ŹRÓDŁOWY DLA OSOBY EDYTUJĄCEJ ARTYKUŁ, Z PRAWEJ – WERSJA DLA CZYTELNIKÓW WIKIPEDII)**



Źródło: opracowanie własne na podstawie danych z Wikipedii.

Zmiany wprowadzane przez użytkowników do każdego artykułu są zapisywane w historii edycji, która pozwala na śledzenie zmian i umożliwia przywrócenie zawartości do poprzedniej wersji. Do sierpnia 2018 użytkownicy dokonali łącznie ponad 2,4 miliarda edycji we wszystkich wersjach językowych [**196**].

W odróżnieniu od innych popularnych serwisów internetowych, Wikipedia nie wyświetla reklam i utrzymuje się z darowizn od użytkowników. Według niektórych szacunków Wikipedia mogłaby zarobić na reklamie ponad 2300 mln USD rocznie [**83**]. W 2016 roku koszty utrzymania tej encyklopedii wyniosły około 66 mln USD, podczas gdy przychód z darowizn wyniósł ponad 77 mln USD [**193**].

Pomimo niekomercyjnego charakteru Wikipedii informacje zawarte w tej bazie wiedzy tworzonej przez społeczność mogą wpływać na decyzje biznesowe i konsumenckie. Strony Wikipedii o znanych osobach, firmach, produktach często pojawiają się jako pierwsze w wynikach wyszukiwania Google, Bing, Yandex i innych popularnych serwisów.

Artykuły o określonych produktach mogą powstawać niezależnie w każdej wersji językowej Wikipedii. W związku z tym jakość informacji o tym samym produkcie może się różnić w zależności od języka językami. Należy także zaznaczyć, iż opis produktu w jednej wersji językowej Wikipedii nie musi być zgodny z informacją zapisaną w innym języku.

Często w artykułach Wikipedii umieszcza się wyróżnioną ramkę, która w przejrzysty sposób ma prezentować najważniejsze informacje o podmiocie artykułu, tzw. infoboks. Czytelnicy korzystają z tych ramek, aby uzyskać najważniejsze fakty o produkcie bez analizy treści całego artykułu.

Od strony technicznej infoboks jest szablonem definiowanym przez użytkowników Wikipedii, a następnie, odpowiednio wypełniony, umieszczany w artykule. To pozwala zachować spójny wygląd infoboksów opisujących różne tematy w poszczególnych wersjach językowych Wikipedii. Zmiana w kodzie szablonu automatycznie pociąga za sobą zmiany we wszystkich artykułach z niego korzystających.

Szablon infoboksu składa się z elementów dwóch rodzajów: parametru i jego wartości. Infoboks na każdy temat ma ściśle określony zestaw parametrów, których używa się do opisu określonego podmiotu czy wydarzenia. Na przykład, w infoboksie opisującym samochody można użyć parametru „zbiornik paliwa", natomiast w opisie telefonów komórkowych ten parametr zostanie zignorowany i nie zostanie wyświetlony czytelnikom Wikipedii. Na rys. **4.3** przedstawiony został przykład wypełnionego infoboksu o samochodzie.

Wstawianie infoboksu do artykułu nie jest obowiązkowe, ale zalecane dla stron na określone tematy, dla których infoboksy zostały przewidziane. Parametry infoboksów i ich wartości zazwyczaj wprowadzane są przez użytkowników, którzy mają różne doświadczenie i wiedzę na określony temat. Zatem wymagane jest dodatkowe sprawdzenie jakości tych danych przez bardziej doświadczonych redaktorów.

Niektóre wartości parametrów mogą pochodzić z innych źródeł niż kod artykułu, w którym jest umieszczony dany infoboks. Jednym z takich źródeł jest projekt Wikidane[11] (ang. Wikidata), który powstał w 2012 roku. Podobnie jak Wikipedia, baza danych Wikidane jest projek-

11 https://www.wikidata.org/

tem Wikimedia Foundation. Głównym celem tego projektu jest stosowanie umieszczonych w nim danych w artykułach Wikipedii oraz innych projektach – niektóre parametry określonych infoboksów mogą być uzupełniane automatycznie na podstawie bazy danych Wikidane.

Kolejne źródło, z którego mogą być automatycznie wstawiane wartości do infoboksu, to dane tabelaryczne (z ang. Tabular Data). Dane w tym przypadku przechowywane są na oddzielnej, specjalnej stronie Wikipedii. W celu ekstrakcji danych z tej strony, definiuje się szablony, które wstawiane są następnie do infoboksów w miejscu, gdzie należy wpisać wartość określonego parametru. Przy wykonaniu tego szablonu dane zostają ekstrahowane z tej specjalnej strony. Jest to szczególnie wygodne, gdy należy przeprowadzić aktualizacje podobnych typów parametrów (np. liczby ludności) dla infoboksów z artykułów o podobnej tematyce (np. miasta określonego państwa).

Infoboksy na określony temat często mają swoje odpowiedniki w różnych językach. Na przykład infoboks opisujący samochody w polskiej wersji ma nazwę „Auto infobox", jego odpowiednik w niemieckojęzycznej to „Infobox PKW-Modell", a w anglojęzycznej – „Infobox automobile". Różnice pojawiają się w zestawach oraz nazwach parametrów, co może utrudniać porównanie wartości parametrów danego infoboksu pomiędzy różnymi wersjami językowymi tego samego tematu.

Dane z Wikipedii wykorzystane są do ulepszenia wyszukiwania obiektów w Grafie wiedzy Google (opartym na Freebase[12]) i Open Graph Facebooka oraz w automatycznych odpowiedziach, w tym Wolfram Alpha[13], Evi [**176**] i Watson IBM [**52**]. Tagi geograficzne z Wikipedii są również wykorzystywane przez Google Maps [**182**].

4.3.3. DBpedia

DBpedia jest projektem crowdsourcingowym, którego celem jest ekstrahowanie strukturyzowanych treści tworzonych w różnych projektach Wikimedia. Otrzymana informacja przypomina otwarty graf wiedzy, który jest dostępny dla wszystkich w Internecie.

Początkowo DBpedia była projektem, którego celem było wydobycie strukturyzowanych informacji z Wikipedii i udostępnienie tych danych za pośrednictwem Internetu [**22**]. DBpedia odwzorowuje parametry infoboksów na specjalną ontologię i tym samym umożliwia wskazywanie ekwiwalentnych parametrów w różnych językach.

12 http://www.freebase.com
13 https://www.wolframalpha.com/

Warunkiem wstępnym jest poprawny opis reguł mapowania każdej nazwy parametrów w każdej wersji językowej. Na przykład, dla samochodów parametry „producent" w polskiej wersji, „Marke" w niemieckiej, „manufacturer" w angielskiej będą mapowane do wspólnego „manufacturer" w ontologii DBpedii. Ta ontologia posiada ponad 300 klas z ponad 1600 różnymi opisanymi właściwościami [**95**].

Przy pomocy ontologii i systemu mapowań[14], DBpedia umożliwia unifikację nazw parametrów infoboksów na podobny temat w różnych wersjach językowych. To pozwala na późniejsze porównywanie wartości parametrów w różnych wersjach językowych (patrz rys. **4.4**).

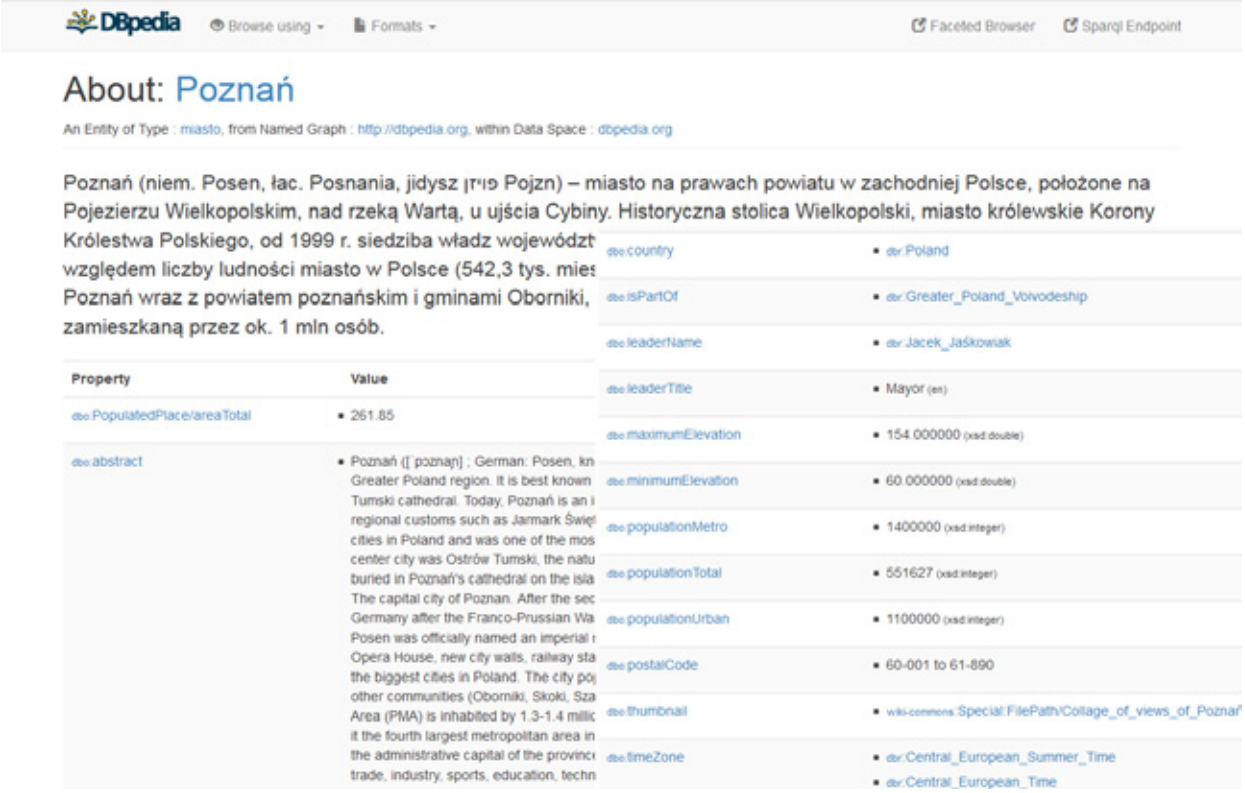
Istnieją badania, które pokazują sposób automatycznej aktualizacji danych w infoboksach na podstawie danych z różnych baz semantycznych, takich jak DBpedia [**3**] czy Wikidata [**149**]. Na rys. **4.5** pokazany jest przykład opisu Poznania w DBpedii z uwzględnieniem danych o liczbie ludności.

RYSUNEK 4.4. **INFOBOKS O MIEŚCIE BAZYLEA Z JEJ ŹRÓDŁAMI DANYCH ORAZ EKSTRAKCJI DANYCH DO DBPEDII Z RÓŻNYCH WERSJI JĘZYKOWYCH WIKIPEDII**

Źródło: opracowanie własne.

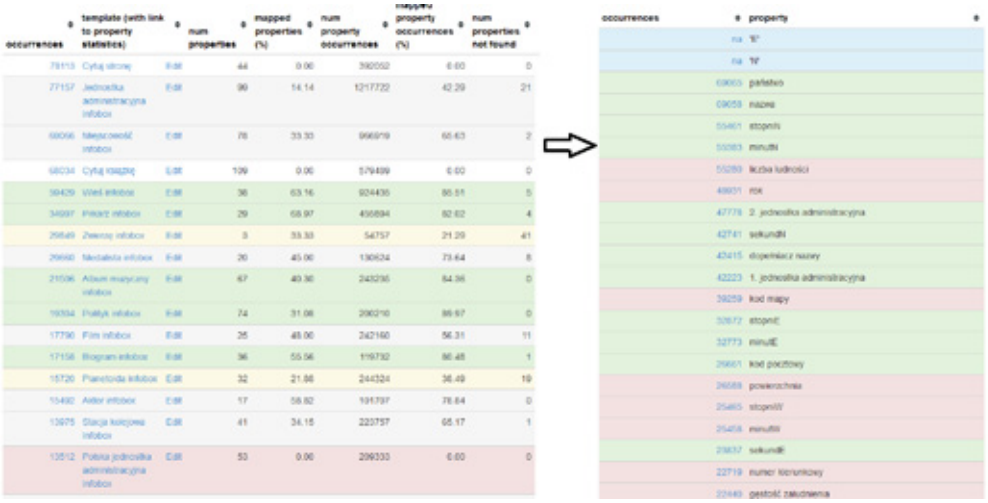
14 http://mappings.dbpedia.org

RYSUNEK 4.5. **STRONA O POZNANIU W DBPEDII Z UWZGLĘDNIENIEM DANYCH O LICZBIE LUDNOŚCI**



Źródło: http://dbpedia.org/page/Pozna%C5%84

RYSUNEK 4.6. **PRZYKŁAD STATYSTYK DBPEDII DOTYCZĄCE MAPOWAŃ POPULARNYCH INFOBOKSÓW POLSKOJĘZYCZNEJ WIKIPEDII ORAZ PARAMETRÓW INFOBOKSU OPISUJĄCEGO MIEJSCOWOŚCI**



Źródło: http://mappings.dbpedia.org/server/statistics/pl/?show=100

Reguły mapowań parametrów infoboksów do cech w DBpedii są tworzone przez użytkowników, jednak istnieją też metody automatyzacji tego procesu [**10**]. DBpedia prowadzi statystyki pokazujące kompletność reguł do mapowania każdego typu infoboksu oraz częstości występowania poszczególnych parametrów każdego infoboksu. Przykład statystyk DBpedii dotyczące mapowań popularnych infoboksów polskojęzycznej Wikipedii oraz parametrów infoboksu opisującego miejscowości przedstawiony został na rys. **4.6**.

Analizując najczęściej wypełniane atrybuty infoboksów wraz z technikami mapowania DBpedii, można zaobserwować różną „kulturę" wprowadzenia danych pomiędzy wersjami językowymi Wikipedii. Rys. **4.7** przedstawia porównanie parametrów infoboksów o grach wideo w różnych wersjach językowych Wikipedii używając mapowań DBpedii.

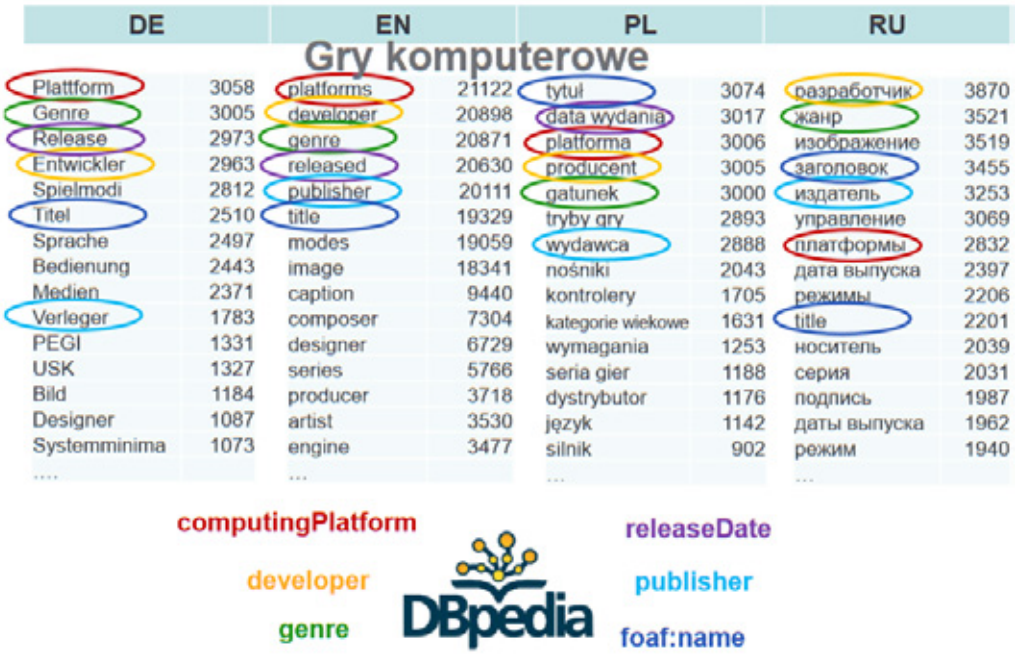
DBpedia zawiera informacje o lokalizacjach i jest powiązana z innymi źródłami danych geograficznych, takimi jak Geonames[15], US Census[16], EuroStat[17], światowy podręcznik faktów CIA[18]

15 https://www.geonames.org/
16 https://www.census.gov/
17 ec.europa.eu/eurostat
18 https://www.cia.gov/library/publications/the-world-factbook/

oraz innych. Zbiór danych zawiera współrzędne geograficzne dla wielu lokalizacji, które umożliwiają wyszukiwanie przy pomocy zapytań w języku SPARQL w zależności od miejsca, w którym jesteśmy. To sprawia, że DBpedia może być cennym źródłem danych dla aplikacji opartych na lokalizacji. Na przykład, DBpedia zawiera krótkie streszczenia o miejscach, które mogą się wyświetlać się na urządzeniach mobilnych [**17**].

Istnieje również możliwość integracji danych z DBpedii do własnej strony WWW. Projekt DBpedia Spotlight[19] może pomóc w ekstrakcji nazwanego obiektu z tekstu, w tym wykrywanie jednostek i rozpoznawanie nazw. Może być również używany do rozpoznawania nazwanych obiektów, wśród innych zadań związanych z wyodrębnianiem informacji. Na rys. **4.8** przedstawiona wersja demonstracyjna systemu DBpedia Spotlight. Serwis posiada również interfejs API[20].

RYSUNEK 4.7. **UNIFIKACJA ATRYBUTÓW INFOBOKSÓW O GRACH WIDEO W RÓŻNYCH WERSJACH JĘZYKOWYCH WIKIPEDII**



Źródło: [103]

19 https://www.dbpedia-spotlight.org/
20 https://www.dbpedia-spotlight.org/api

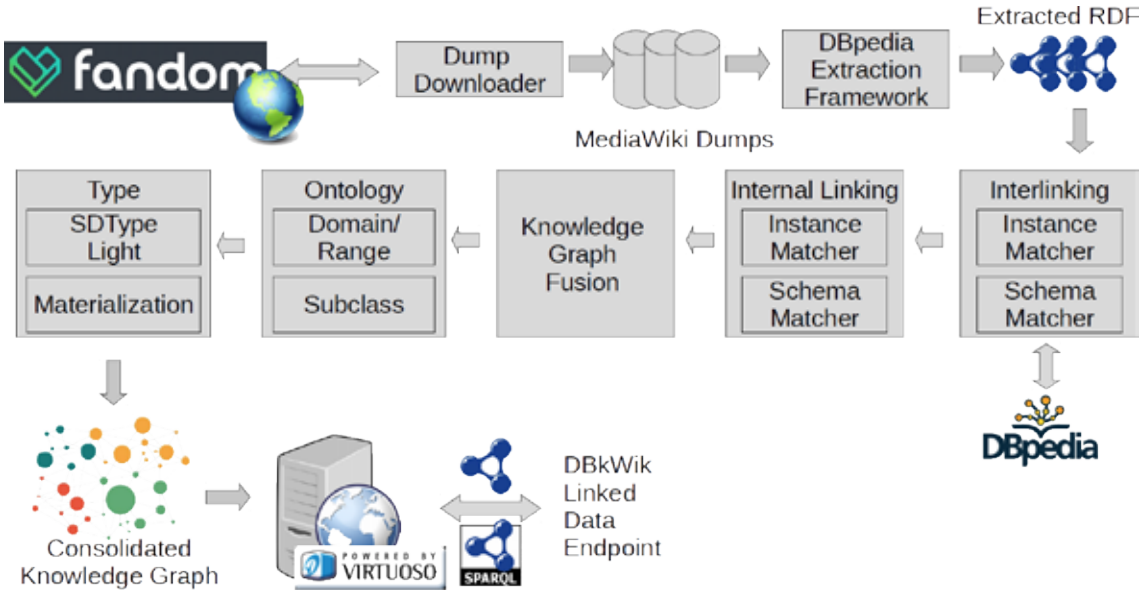
Metody używane w DBpedii mogą pomóc w ekstrakcji informacji nie tylko z Wikipedii. Projekt DBkWik[21] pozwala na ekstrakcję danych z tysięcy serwisów typu wiki, tworząc skonsolidowaną bazę wiedzy [**71**]. Schemat ekstrakcji danych z serwisów wiki pokazany na rys. **4.9**.

RYSUNEK 4.8. **STRONA DEMONSTRACJI SERWISU DBPEDIA SPOTLIGHT**

Źródło: https://www.dbpedia-spotlight.org/demo/

21 http://dbkwik.webdatacommons.org/

RYSUNEK 4.9. **DBKWIK: SCHEMAT EKSTRAKCJI DANYCH Z SERWISÓW WIKI**



Źródło: [71]

Narzędzie do ekstrakcji danych DBpedia Ekstraction Framework można uruchomić na własnym komputerze. Przed rozpoczęciem ekstrakcji musimy zadbać o instalacje wszystkich niezbędnych komponentów opisanych na stronie[22], m.in. to:

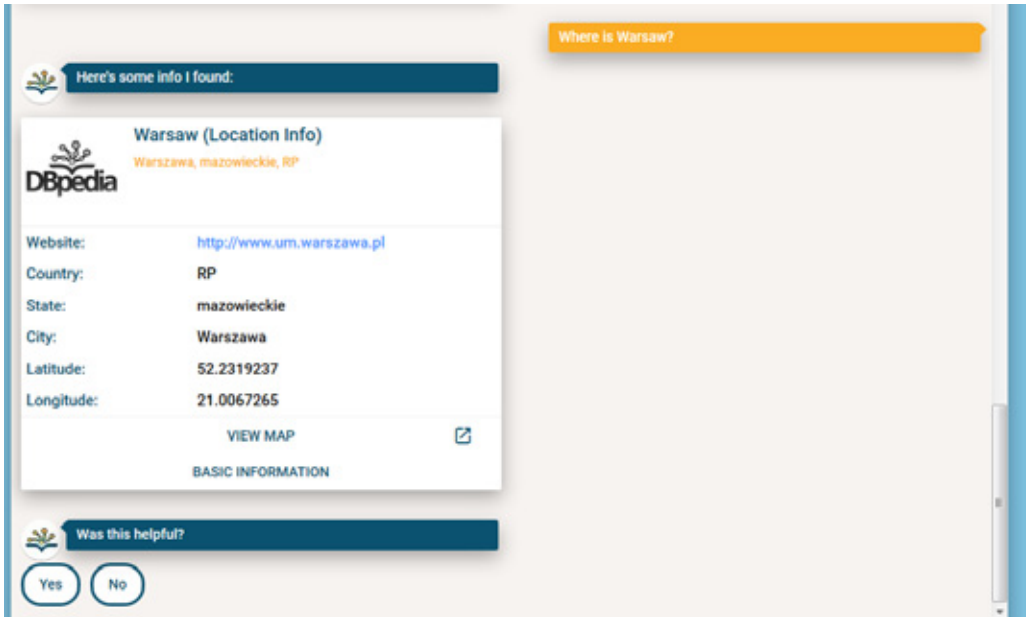
- DBpedia Extraction Framework - kod zródłowy na GitHub[23].
- Java Development Kit[24] – do działania ekstraktor wymaga Java co najmniej wersji 7.
- Maven[25] – służy do zarządzania projektami i automatyzacji kompilacji.

Do pracy DBpedia Extraction Framework wymaga kopie zapasowe Wikipedii w wybranych wersjach językowych, które można pobrać na z odpowiedniej strony fundacji Wikimedia[26]. Należy również wybrać, z których ekstraktorów chcemy korzystać, m.in. to[27]:

22 https://github.com/dbpedia/extractionframework/wiki/Documentation
23 https://github.com/dbpedia/extraction-framework
24 https://www.oracle.com/technetwork/java/ javase/downloads/index.html
25 http://maven.apache.org/
26 https://dumps.wikimedia.org/backup-index.html
27 Pełna lista ekstraktorów DBpedia Ectraction Framework z opisem dostępna na stronie https://wiki.dbpedia.org/services-resources/documentation/extractor

- LabelExtractor – wyodrębnia etykiety do artykułów na podstawie ich tytułu.
- MappingExtractor – wyodrębnia dane strukturalne w oparciu o ręcznie generowane odwzorowania (mapowania) infoboksów Wikipedii do ontologii DBpedia. Odwzorowania można edytować za pomocą Mappings Wiki[28].
- InfoboxExtractor – ekstrahuje wszystkie właściwości ze wszystkich infoboksów. Wyodrębnione informacje są reprezentowane za pomocą właściwości w obszarze nazw http://dbpedia.org/property/.
- WikiPageExtractor – wyodrębnia linki do odpowiednich artykułów w Wikipedii.
- PageLinksExtractor – ekstraktacja linki wewnętrznych między instancjami DBpedii z wewnętrznych linków artykułów Wikipedii. Linki do stron mogą być przydatne do analizy strukturalnej, eksploracji danych lub do klasyfikowania instancji DBpedia przy użyciu Page Rank lub podobnych algorytmów.
- GeoExtractor – odpowiada za ekstrakcję współrzędnych geograficznych.
- CategoryLabelExtractor – wyodrębnia etykiety dla kategorii.
- ImageExtractor – Wyodrębnia pierwszy obraz strony Wikipedii. Tworzy miniaturę oraz obraz w pełnym wymiarze.

RYSUNEK 4.10. **DBPEDIA BOT: PRZYKŁAD ODPOWIEDZI NA PYTANIE**



Źródło: http://chat.dbpedia.org/

28 http://mappings.dbpedia.org/index.php/Main_Page

Przy pomocy danych z DBpedii można tworzyć interaktywne aplikacje, takie jak np. chat boty. Jedna z wersji takiego bota dostępna jest pod adresem: http://chat.dbpedia.org/. Przykład pytania oraz odpowiedzi w ramach tego narzędzia pokazany na rys. **4.10**.

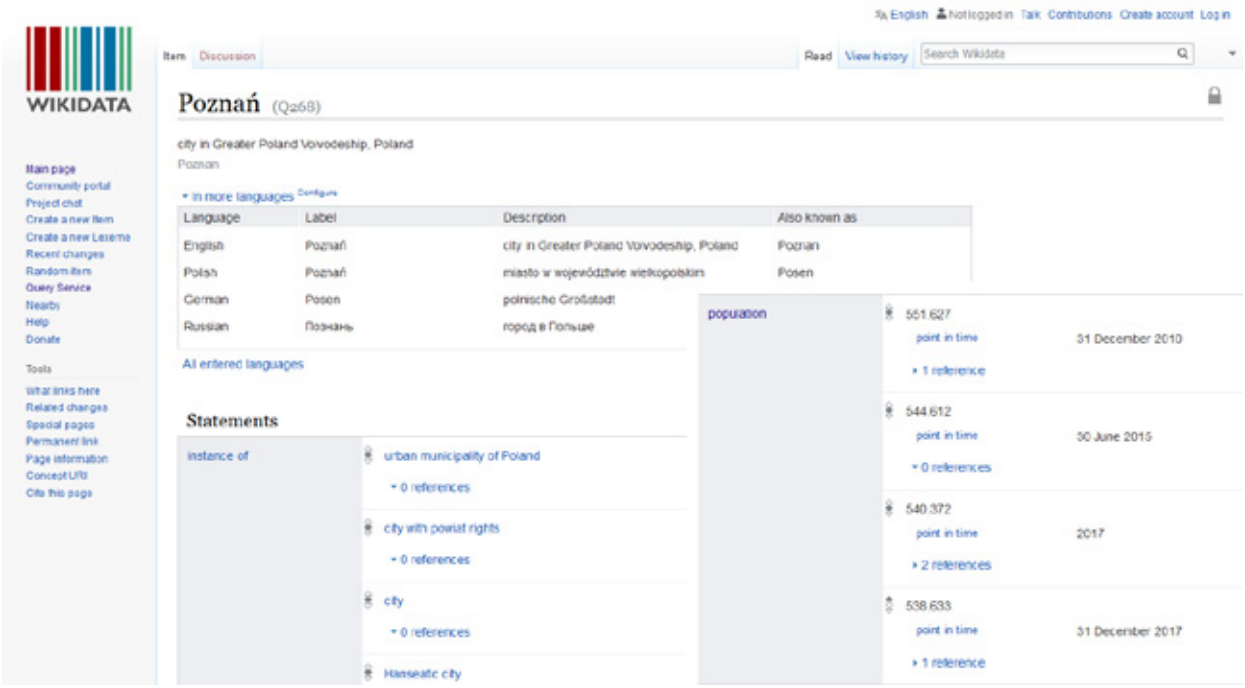
4.3.4. Wikidane

Wikidane[29] – to bezpłatna baza wiedzy, którą mogą czytać i edytować: zarówno ludzie, jak i maszyny. Wikidane umożliwiając globalnej społeczności edycję danych, które mogą pochodzić i innych zbiorowych źródeł [**182**]. Ta semantyczna baza obsługuje użycie dowolnych właściwości we wszystkich obiektach, których obecnie jest tam umieszczonych ponad 54 mln[30]. Dane można umieszczać indywidualnie przez specjalny interfejs lub wsadowo przy pomocy importowania zestawu danych[31].

Aby przechowywać dane strukturalne poza etykietami tekstowymi i linkami językowymi, Wikidane używają prostego modelu danych. Dane są zasadniczo opisane za pomocą par własności i wartości; na przykład pozycja „Poznań" może mieć właściwość „population" (liczba ludności) o wartości 538.633. Właściwości to dodatkowo obiekty, które mają własne strony Wikidata z etykietami, aliasami i opisami. W przeciwieństwie do przedmiotów, strony te nie są powiązane z artykułami Wikipedii. Na rys. **4.11** pokazany przykład opisu Poznania w Wikidanych z uwzględnieniem danych o populacji. Należy zwrócić uwagę, że dla jednej cechy (w danym przykładzie „population") może mieć różne wartości, które mogą wynikać m.in. z różnych źródeł oraz różnic pomiędzy wersjami językowymi artykułów Wikipedii, z których te dane pochodzą.

29 https://www.wikidata.org/wiki/Wikidata:Main_Page
30 https://www.wikidata.org/wiki/Special:Statistics
31 https://www.wikidata.org/wiki/Wikidata:Dataset_Imports

RYSUNEK 4.11. **STRONA O POZNANIU W WIKIDANYCH Z UWZGLĘDNIENIEM DANYCH O POPULACJI**



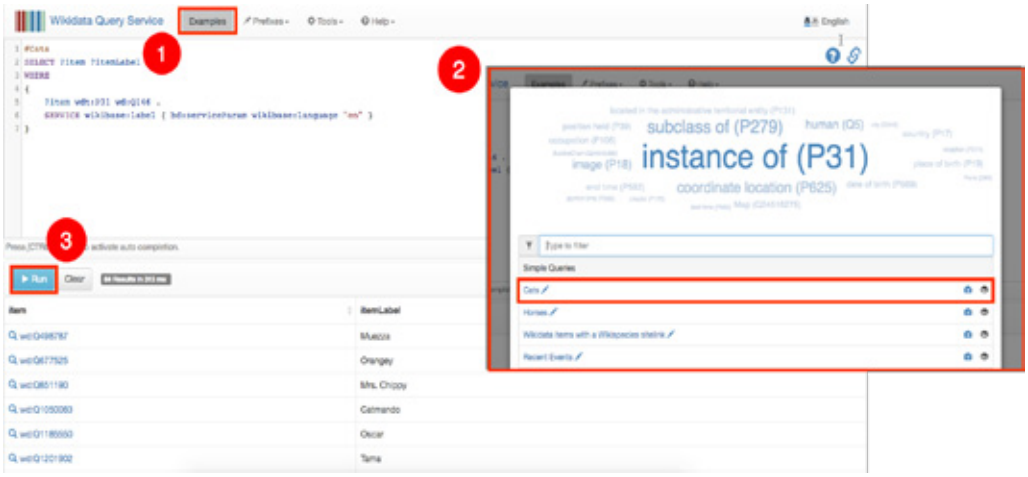
Źródło: https://www.wikidata.org/wiki/Q268

Wikidane umożliwiają robić zapytania online przy pomocy języka SPARQL na stronie https://query.wikidata.org/. Można skorzystać z listy przykładów w tym narzędziu. Do tego należy kliknąć na „Examples" („Przykłady"), w pojawiającym się oknie wybrać z listy interesujący przykład (np. „Cats"), i na końcu wciągnąć „Run" (Uruchom) tuż pod polem kodu. Schemat postępowania pokazany na rys. **4.12**.

Istnieje wiele sposobów wykorzystania Wikidanych w innych projektach Wikimedia:

- Wikidane wzbogacają infoboksy na Wikipedii,
- Mapy i wykresy,
- Wizualizacje.

RYSUNEK 4.12. **SCHEMAT POSTĘPOWANIA DO URUCHOMIENIA PRZYKŁADOWEGO ZAPYTANIA W WIKIDATA SPARQL**



Źródło: https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/A_ gentle_introduction_to_the_Wikidata_Query_Service

4.3.5. Common Crawl

Common Crawl to inicjatywa, której celem jest demokratyzacja dostępu do informacji sieciowych poprzez tworzenie i utrzymywanie otwartego repozytorium danych indeksowania stron internetowych, które są powszechnie dostępne do odczytu i analizy. Ta inicjatywa jest wspierana przez organizację non-profit Common Crawl Foundation z siedzibą w USA.

Common Crawl robi to, co robią Google, Bing oraz inne wyszukiwarki, ale pozwala każdemu uzyskać dostęp do swoich danych i analizować ich za darmo z możliwością wykorzystania komercyjnego.

Zebrane dane przez Common Crawl są zindeksowane w taki sposób, że umożliwią obserwowanie ewolucji podstawowej struktury sieci WWW w ciągu ostatnich 10 lat [**117**]. Archiwum indeksowania stron tworzony jest co miesiąc [**34**]. W grudniu 2018 roku Common Crawl zawierał około 3.1 mld stron WWW całkowitą objętością około 250 TiB nieskompresowanej zawartości [**156**].

Robot używany do indeksowania w celu przeszukiwania bazuje na strategii przeszukiwania wszerz, wraz z heurystyką wykrywania stron ze spamem. Ponadto zastosowano heurystyki w celu zmniejszenia liczby przeszukiwanych stron z duplikatem lub bez zawartości [**117**].

Do zbierania danych Common Crawl używa własne algorytmy, podobne do robotów internetowych znanych wyszukiwarek, w związku z tym uwzględniane są możliwe ograniczenia stron

WWW. Niektóre z tych ograniczeń wystawiają właściciele witryn przy pomocy pliku „robots.txt", aby przekazać instrukcje dotyczące robotom. Te instrukcji mogą zawierać m.in. listę plików lub folderów, których odwiedzanie oraz indeksowanie jest zabronione. Przy tym, te ograniczenia mogą dotyczyć wszystkich badź wybranych wyszukiwarek. Działa to w następujący sposób: robot chce odwiedzić URL strony internetowej, np. http://kie.ue.poznan.pl/pl/members. Zanim to zrobi, najpierw sprawdza http://kie.ue.poznan.pl/robots.txt i znajduje:

```
1  User-agent: *
2  Disallow: /
```

„User-agent: *" oznacza, że ta sekcja dotyczy wszystkich robotów. „Disallow: /" mówi robotowi, że nie powinien odwiedzać żadnych stron w witrynie.

CommonCrawl dostarcza dane w 3 formatach:

- PlikiWARCprzechowującesurowedaneindeksowania.Dlawersjizestycznia2019 roku to są dane objętością około 59 TB po skompresowaniu.
- Pliki WAT przechowujące wyekstrahowane metadane dla danych przechowywanych w WARC. Dla wersji ze stycznia 2019 roku to są dane objętością około 19 TB po skompresowaniu.
- Pliki WET, które przechowują wyodrębniony tekst z danych przechowywanych w WARC. Dla wersji ze stycznia 2019 roku to są dane objętością około 9 TB po skompresowaniu.

4.3.6. Web Data Commons

Coraz więcej witryn internetowych umieszcza dane strukturalne opisujące osoby, produkty, organizacje, wydarzenia, recenzje, wpisy na blogu, i nawet przepisy kulinarne na swoich stronach HTML przy użyciu takich stosunkowo nowych standardów jak Mikroformaty, Mikrodane oraz RDF [**23**] Te standardy pozwalają na wzbogacenie informacji na stronach o dodatkowe metadane. Takie podejście pozwala oprogramowaniu automatycznie przetwarzać informacje przeznaczone dla użytkowników końcowych (np. współrzędne geograficzne, terminy wydarzeń, dane kontaktowe i inne).

Projekt Web Data Commons[32] (WDC) wyodrębnia te dane z kilku miliardów stron internetowych przy użyciu Common Crawl, który był opisany w poprzednim podrozdziale. Projekt dostarcza wyodrębnione dane do pobrania i publikuje statystyki dotyczące rozmieszczania różnych formatów. Dane są ekstrahowane co roku.

W wersji za listopad 2018 roku projekt WDC po analizie około 54 TB skompresowanych danych stron internetowych wyekstrahował ponad 31 miliardów semantycznych trójek [**187**].

32 http://webdatacommons.org

4.3.7. OpenStreetMap

Na całym świecie kilkaset tysięcy osób współtworzy informacje geograficzne. Jednym z najbardziej znanych projektów w tym kierunku jest OpenStreetMap (OSM), który był założony w 2004 roku. Ten projekt cieszył się w ostatnich latach wyjątkowym zainteresowaniem i jest uważany za jedno z najbardziej imponujących źródeł informacji geograficznej w Internecie [**130**].

Jedno z najprostszych zastosowań OSM to geolokalizacja oraz szukanie ścieżek dla użytkowników końcowych. Na przykład, do użycia map z OSM na urządzeniach firmy Garmin należy wykonać następujące kroki:

- Pobrać dane OSM.
- Podzielić duży obszar
- Skonwertować do pliku IMG.
- Utworzyć pliki GMAPSUPP.IMG (opcjonalnie)

Jeżeli chodzi o pobieranie danych z OSM, istnieją różne możliwości:

- Najłatwiejszym sposobem jest użycie przycisku „Eksport" („Export") na stronie OSM, następnie w wyświetlonym oknie wybrać „Eksportuj" (dla mniejszych obszarów) lub wybrać „Overpass API". Spowoduje to pobranie mapy, którą aktualnie jest na wyświetlona w przeglądarce.
- Jeśli chcemy pobrać mapę całego kraju lub świata, możemy wykorzystać zrzuty (zajmuje około 1 TB) ze strony https://wiki.openstreetmap.org/wiki/Planet.osm.
- Przy użyciu aplikacji JOSM[33] należy wybrać obszar, następnie pobrać z serwera oraz zapisać do pliku.
- Przy użyciu wiersza poleceń
- Jeżeli chcemy wdrożyć mapy do swojej aplikacji, możemy użyć Overpass QL[34]

W celu uproszczenia czynności związanych z pobieraniem oraz używaniem map OSM, można również skorzystać z różnych narzędzi:

- Mkgmap[35] – program z otwartym kodem źródłowym napisany w języku Java opracowany specjalnie do konwersji danych OSM w pliki obrazów map do urządzeń Garmin. Zaleca się dla początkujących.
- osm2mp[36] serwis do tworzenie map Garmin z trasowaniem konwertowaniem do polskiego formatu map (MP).

33 https://josm.openstreetmap.de/
34 https://wiki.openstreetmap.org/wiki/Overpass_API/Overpass_QL
35 https://www.mkgmap.org.uk
36 https://forum.openstreetmap.org/viewtopic.php?id=1162

- QLandkarte[37] – program z otwartym kodem źródłowym do wyświetlania i pracy z mapami Garmin. Program nie jest już aktualizowany.
- Mapwel[38] – kompleksowy program do mapowania GPS. Obsługuje pliki OSM i konwersję ich do formatu Garmin IMG z trasowaniem.
- OSM Composer[39] – to aplikacja z graficznym interfejsem upraszczająca proces tworzenia obrazów mapy Garmin z danych OSM i SRTM.
- Osm2garminGUI – aplikacja na języku Java, która umożliwia pobieranie danych OSM dla całej planety, aktualizacja istniejących danych planety, dodawanie konturów terenu wygenerowanych z bazy danych SRTM3 i generowanie map zgodnych ze standardem Garmin w jednym kroku.
- Inne[40].

Innym dobrym przykładem serwisu używającego mapy OSM jest MapQuest[41]. W tym serwisie można edytować mapy, zaznaczać trasy oraz ulubione lokalizacje. Istnieje również możliwość osadzać niestandardową mapę MapQuest na własnej witrynie lub aplikacji. Rys. **4.13** ilustruje przykład użycia MapQuest w komunikacji miejskiej w Poznaniu.

RYSUNEK 4.13. **PRZYKŁAD UŻYCIA MAPQUEST: EKRAN ZAMONTOWANY W TRAMWAJU NR 5 W POZNANIU POKAZUJE BIEŻĄCĄ LOKALIZACJĘ**

Źródło: zdjęcie własne.

37 http://qlandkarte.sourceforge.net/
38 http://www.mapwel.net/
39 https://wiki.openstreetmap.org/wiki/OSM_Composer
40 https://wiki.openstreetmap.org/wiki/OSM_Map_On_Garmin
41 https://www.mapquest.com/

4.3.8. Inne bazy społecznościowe

Istnieje wiele innych przykładów baz wiedzy, współtworzonych przez społeczność.

Spośród nich warto wyróżnić:

- WordNet, popularna leksykalna baza danych dla różnych języków.
- VerbNet, największy internetowy leksykon czasownikowy.
- GeoNames, darmowa geograficzna baza danych obejmująca wszystkie kraje i zawierająca ponad osiem milionów nazw miejsc.
- ImageNet, baza danych obrazów zorganizowana zgodnie z hierarchią WordNet (2011 release).

4.4. Agregatory zbiorów danych

W Internecie istnieją specjalne katalogi oraz wyszukiwarki otwartych baz danych, z możliwością dodawania opisu do nowych oraz wprowadzać zmiany do istniejących zasobów.

- OpenML[42] – ponad 2.5 tys. zweryfikowanych zbiorów danych, służących m.in. do budowania modeli z wykorzystaniem algorytmów uczenia maszynowego. OpenML oferuje różne usługi udostępniania i znajdowania zestawów danych, pobierania lub tworzenia zadań naukowych, udostępniania i znajdowania implementacji oraz udostępniania i organizowania wyników [**178**]. Usługi są dodatkowo dostępne za pośrednictwem interfejsu API (zob. sekcja **6.5.2**).
- OpenDOAR[43] – około 3800 repozytoriów, które w większości zakładają się z zdigitalizowanych zbiorów instytucji akademickich (np. biblioteki). Dostępna wyszukiwarka oraz katalog z podziałem na kontynenty i państwa. Posiada również interfejs API, przy pomocy którego można uzyskać meta dane wybranych repozytoriów[44].
- Data.World[45] – ponad 400 opisów różnych otwartych baz danych. Do wyświetlenia szczegółowej informacji obowiązkowa rejestracja na portalu.
- Data Hub Core[46] – dostęp do około 80 otwartych baz danych, które w większości dotyczą wskaźników ekonomicznych. Posiada interfejs API. Dodatkowo, dla każdego zbioru danych serwis przedstawia możliwości wykorzystania danych przy użyciu Python, R, JavaScript, Pandas oraz innych. Istnieje możliwość uzyskania dostępu do szerszej listy baz danych za dodatkową opłatę.

42 https://www.openml.org/
43 http://v2.sherpa.ac.uk/opendoar/
44 http://v2.sherpa.ac.uk/opendoar/api.html
45 https://data.world/datasets/open-data
46 https://datahub.io/core

- AWS Open Data[47] – rejestr otwartych baz danych od firmy Amazon. Zawiera około 100 baz danych, które można uzupełniać.
- Re3data[48] – szczegółowe informacje na temat ponad 2000 repozytoriów danych badawczych.
- Academic Torrents[49] – ponad 2000 otwartych zbiorów danych.
- Data Portals[50] – serwis zawiera opis około 550 portali z danymi.
- OpenDataSoft[51] ponad 17 tys. otwartych baz danych. Posiada interfejs API.
- UCI[52] – 468 otwartych baz danych do uczenia maszynowego.
- Kolekcja dużych zbiorów danych Uniwersytetu Stanforda[53] – zawiera wybrane zbiory danych z sieci społecznościowych (tj. Facebook, Twitter, Google+, LiveJournal), cytowań (np. Arxiv), webowe (np. Google, Stanford), zakupów produktów (np. Amazon), sieci typu peer-to-peer (Gnutella), Wikipedii, recenzje (np. Amazon, CellarTracker) oraz inne
- Figshare[54] – to internetowe repozytorium otwartego dostępu, w którym można przechowywać oraz dzielić się wynikami badań, w tym zestawami danych, obrazami i filmami (zob. sekcja **6.5.6**).

W niektórych popularnych serwisach internetowych, można znaleźć specjalne strony (profili lub kategorie), w których również gromadzone bazy lub odnośniki do nich.

Niżej przedstawiona lista takich stron na popularnych serwisach internetowych:

- https://www.reddit.com/r/datasets/ – kategoria zawierająca linki do różnorodnych baz danych, umieszczona w ramach serwisu społecznego Reddit. Serwis ten umożliwia umieszczanie treści i linków do różnorodnych informacji w Internecie, z ich jednoczesnym grupowaniem. Wszystkie treści można oceniać oraz komentować. Domyślnie linki są uporządkowane według czasu publikacji (od najnajnowszego), jednak na miejsce wpisu na liście mogą wpłynąć oceny użytkowników. Do analizy najwyżej ocenianych treści można korzystać z dodatkowych narzędzi, takich jak http://redditmetrics.com/.
- https://en.wikipedia.org/wiki/List_of_datasets_for_ machine_learning_research – strona na Wikipedii, służąca do agregacji różnych otwartych baz danych ze wskazaniem liczby rekordów, ostatniej aktualizacji, autora oraz źródła. Dodatkowo zbiory danych zostały

47 https://registry.opendata.aws/
48 https://www.re3data.org
49 http://academictorrents.com/
50 http://dataportals.org/
51 https://data.opendatasoft.com/pages/home/
52 https://archive.ics.uci.edu/ml/datasets.html
53 https://snap.stanford.edu/data/
54 https://figshare.com

pogrupowanie według tematów: obrazki, dane tekstowe, dźwiękowe, biologiczne, fizyczne, wielowymiarowe (takie jak pogoda, gry, spis ludności) oraz inne.

- Na serwisie GitHub można znaleźć kilka stron z regularnie sprawdzaną oraz aktualizowaną listą potencjalnie przydatnych baz danych, m.in. to:
  – https://github.com/openml/OpenML/wiki/Data-Repositories,
  – https://github.com/awesomedata/awesome-publicdatasets/,
  – https://github.com/niderhoff/nlp-datasets.

4.5. Zbiory do głębokiego uczenia

Cechą charakterystyczną uczenia maszynowego, a w szczególności głębokiego uczenia, jest zapotrzebowanie na wysokiej jakości dane. Zdolności predykcyjne modeli zależą od jakości danych wejściowych – niezależnie od poziomu zaawansowania architektury. Firmy, uniwersytety i inne jednostki badawcze coraz częściej decydują się na publiczne udostępnianie baz wykorzystywanych przy własnych badaniach. Pozwala to tworzenie i testowanie modeli w oparciu o drobiazgowo przygotowane i wysokiej jakości dane wejściowe bez potrzeby samodzielnego ich gromadzenia (który to proces może być czasochłonny i kosztowny).

4.5.1. Analiza obrazów

- **LabelMe**[55]

Projekt prowadzony przez Computer Science and Artificial Intelligence Laboratory przy Massachusetts Institute of Technology. Jego celem było zbudowanie i publiczne udostępnienie zbioru danych, pozwalającego na trenowanie modeli do rozpoznawania, wykrywania i lokalizowania obiektów na zdjęciach.

LabelMe skupia się na złożonych scenach. Zamiast dostarczać zdjęcia przygotowane tak, by przedstawiały jeden konkretny obiekt, wykorzystano wielokątne ramki, nanoszone przez ludzi, w celu wyróżnienia wszystkich instancji obiektów, które pojawiły się na danej fotografii. Niektóre zawierają do kilkuset oznaczonych i opisanych elementów. Obecnie LabelMe liczy 178371 zdjęć.

Istotną częścią projektu jest aplikacja LabelMe Annotation Tool, która pozwala na opisywanie własnych zdjęć i dodawanie ich do publicznego repozytorium. Korzystać może z niego każdy, po założeniu konta na stronie internetowej. Kod źródłowy aplikacji znajduje się w domenie publicznej na licencji MIT, dostępny jest w serwisie Github[56]. Zgodnie z licencją, dodane do bazy zdjęcia przechodzą do domeny publicznej. Jest to dobre rozwiązanie dla korzystających z takich zdjęć.

55 http://labelme2.csail.mit.edu/Release3.0/ browserTools/php/dataset.php
56 https://github.com/CSAILVision/LabelMeAnnotationTool

- **ImageNet**[57]

Najbardziej znany i najpowszechniej wykorzystywany przez badaczy zbiór danych dotyczący rozpoznawania i wykrywania obiektów na zdjęciach. Tworzony przez uniwersytety Stanford oraz Princeton. Zorganizowany jest zgodnie z hierarchią WordNet[58], wykorzystując jako nazwy klas pochodzące z niego zbiory synonimów (ang. *synsets*). ImageNet zawiera ponad 14 milionów zdjęć, podzielonych na niemal 22 tysiące kategorii. Średnio każda z kategorii zwiera 500 zdjęć. Wszystkie obrazy zostały opisane przez ludzi, co jest przesłanką do wysokiej oceny jakości danych. ImageNet jako taki nie dostarcza zdjęć, a tylko adresy URL do nich oraz miniaturki i opisy. Prawa autorskie do zdjęć posiadają ich autorzy.

Istotną częścią projektu ImageNet jest konkurs, Large Scale Visual Recognition Challenge (LSVRC), organizowany corocznie od roku 2010. Biorą w nim udział zespoły badaczy z całego świata, prezentując modele, które mają osiągnąć jak największą dokładność w zadaniach wykrywania oraz lokalizowania obiektów. Rywalizacja ta doprowadziła do rozpowszechnienia wielu przełomowych rozwiązań w dziedzinie automatycznego przetwarzania obrazów – najbardziej znanym przykładem jest LSVRC 2012, kiedy po raz pierwszy konkurs wygrała konwolucyjna sieć neuronowa. Od tamtej pory architektury tego rodzaju stały się wiodącym standardem.

RYSUNEK 4.14. **PRZYKŁAD HIERARCHICZNEJ ORGANIZACJI OBIEKTÓW W ZBIORZE IMAGENET**

Źródło: [44]

57 http://image-net.org
58 https://wordnet.princeton.edu/

- **COCO – Common Objects in Context**[59]

Stworzony i opublikowany w 2014 roku przez Microsoft zbiór danych, stawiający sobie podobne cele, co LabelMe. COCO skupia się na problemie przetwarzania i zrozumienia złożonych scen, przedstawiających rozmaite obiekty w ich naturalnym otoczeniu. Łącznie zawiera ponad 330 tysięcy obrazów. Dzieli się na pięć części, dotyczących różnych zagadnień:

- wykrywanie obiektów z wykorzystaniem technik ramek granicznych oraz segmentacji obiektów. Zawiera ponad 200 tysięcy zdjęć oraz 80 kategorii obiektów,
- lokalizacja punktów kluczowych postaci ludzkich. Zawiera ponad 200 tysięcy zdjęć, na których znajduje się 250 tysięcy sylwetek ludzkich z opisanymi punktami kluczowymi,
- *stuff detection*, kategoria skupiająca się na wykrywaniu elementów zdjęć, które nie posiadają zwykle określonych kształtów i wymiarów (jak np. samochód lub pies), a które rozpoznaje się raczej na podstawie powtarzających się wzorców i struktury (np. niebo, trawa, ściana). Wykrywanie takich obiektów pozwala na lepsze opisywanie scen zawartych na zdjęciach. Podzbiór liczy 163 tysiące zdjęć podzielonych na 92 kategorie.
- wizualna segmentacja scen, kategoria dotycząca opisywania wszystkiego, co znajduje się na zdjęciu, łączy w sobie wykrywanie oraz lokalizację obiektów oraz elementów tła. Wykorzystuje cały zbiór COCO.
- generowanie złożonych podpisów do zdjęć, opisujących zawarte na nich obiekty, elementy tła, relacje między nimi oraz czynności.

Ze względu na wysoką jakość opisów oraz wsparcie dla wielu różnych rodzajów zdań, COCO cały czas zyskuje na popularności wśród badaczy. Na koncie projektu w serwisie Github można znaleźć między innymi API[60]. Prawa do opisów zdjęć udostępnione zostały na licencji Creative Commons Attribution 4.0. Prawa do zdjęć stanowią własność ich autorów.

- **COIL-100**[61]

Zbiór stworzony przez Computer Vision Laboratory, działające przy Uniwersytecie Columbia. Zawiera 7200 zdjęć przedstawiających 100 obiektów na czarnym tle. Każdy z obiektów został sfotografowany 72 razy – przed wykonaniem każdego kolejnego zdjęcia obiekt obracany był o 5 stopni, dlatego 72 fotografie pozwalają przedstawić go pod każdym możliwym kątem podczas obrotu o 360 stopni. Zdjęcia przedstawiają przedmioty codziennego użytku (zabawki, elementy wyposażenia domowego), warzywa oraz owoce i opakowania produktów (puszki, pojemniki).

59 http://cocodataset.org/
60 https://github.com/ cocodataset
61 http://www1.cs.columbia.edu/CAVE/software/softlib/coil100.php

- **Labeled Faces in the Wild**[62]

Stworzony przez badaczy z Uniwersytetu Massachusetts, skupia się na zagadnieniu rozpoznawania twarzy. Zawiera 13233 zdjęcia przedstawiające 5749 różnych osób. Każde ze zdjęć podpisane jest nazwiskiem. W oryginalnej formie nadaje się przede wszystkim do trenowania modeli rozpoznających konkretne osoby. Po odpowiednim opisaniu mógłby również zostać wykorzystany np. do analizy wydźwięku lub rozpoznawania emocji czy punktów charakterystycznych twarzy.

- **Google Open Images**[63]

Opublikowany przez Google zbiór, zawierający około 9 milionów zdjęć. Wszystkie posiadają ogólne opisy, wygenerowane automatycznie, z których część została sprawdzona przez ludzi. Ponadto zbiór zawiera ponad 1,7 miliona zdjęć, na których oznaczono obiekty za pomocą ramek. Łącznie zaznaczono ponad 14,6 miliona obiektów podzielonych na 600 klas. Średnio zaznaczonych obiektów to 8,4. To największy obecnie zbiór tego typu. Wszystkie ramki zostały zaznaczone przez ludzi, co jest przesłanką do wysokiej oceny jakości danych. Prawa do opisów zdjęć udostępnione zostały na licencji Creative Commons Attribution 4.0, a zdjęcia Creative Commons Attribution 2.0 (autorzy zbioru zastrzegają jednak obowiązek samodzielnej weryfikacji praw do zdjęć).

- **Indoor scene recognition**[64]

Stworzony przez badaczy z Massachusetts Institute of Technology zbiór danych poświęcony rozpoznawaniu i klasyfikowaniu wnętrz budynków. Zgromadzone zdjęcia biorą pod uwagę główny problem tego rodzaju klasyfikacji – wykorzystanie nie tylko właściwości przestrzennych pomieszczeń, ale również przedmiotów, które się w nich znajdują, odpowiednio łącząc oba źródła informacji. Opublikowana baza zawiera 15620 zdjęć, podzielonych na 62 kategorie (np. sklepy, biura czy przestrzenie użyteczności publicznej takie jak lotniska i biblioteki). Każda z kategorii zawiera przynajmniej 100 zdjęć. Zbiór został opublikowany wyłącznie dla celów badawczych.

- **Flickr 8K**[65]

Opracowany przez badaczy z Uniwersytetu Illinois zbiór 8092 zdjęć przeznaczonych do budowania modeli generujących złożone opisy scen. Zdjęcia przedstawiają ludzi i zwierzęta (głównie psy). Pobrane zostały z portalu społecznościowego Flickr, skupiającego amatorów fotografii. Następnie, za pośrednictwem Amazon Mechanical Turk[66], każdy z obrazów został opa-

62 http://vis-www.cs.umass.edu/lfw/
63 https://storage.googleapis.com/openimages/web/index.html
64 http://web.mit.edu/torralba/www/indoor.html
65 https://forms.illinois.edu/sec/1713398
66 https://www.mturk.com/

trzony przez ludzi pięcioma opisami, które miały jak najdokładniej opisać jego zawartość. Dane udostępniane są przez autorów dla celów badawczych i edukacyjnych. Autorzy zastrzegają, że nie posiadają praw autorskich do zdjęć.

4.5.2. Analiza wydźwięku

- **IMDB reviews**[67]

Opracowany na Uniwersytecie Stanforda zbiór danych pozwalający na trenowanie modeli do binarnej klasyfikacji wydźwięku. Zawiera łącznie 50000 recenzji filmów, zamieszczonych przez użytkowników portalu IMDB[68], sklasyfikowanych jako pozytywne lub negatywne. Obie grupy zawierają po 25000 recenzji. Baza posiada również dodatkowy zestaw niesklasyfikowanych danych (50000 recenzji), który może zostać wykorzystany w procesie nienadzorowanego uczenia maszynowego. Pliki zawierają również przyznaną przez użytkownika ocenę w dziesięciostopniowej skali. Dane dostępne są w dwóch formatach – tekstowym oraz w modelu Bag of Words, powszechnie wykorzystywanym do przetwarzania języka naturalnego.

- **Stanford Sentiment Treebank**[69]

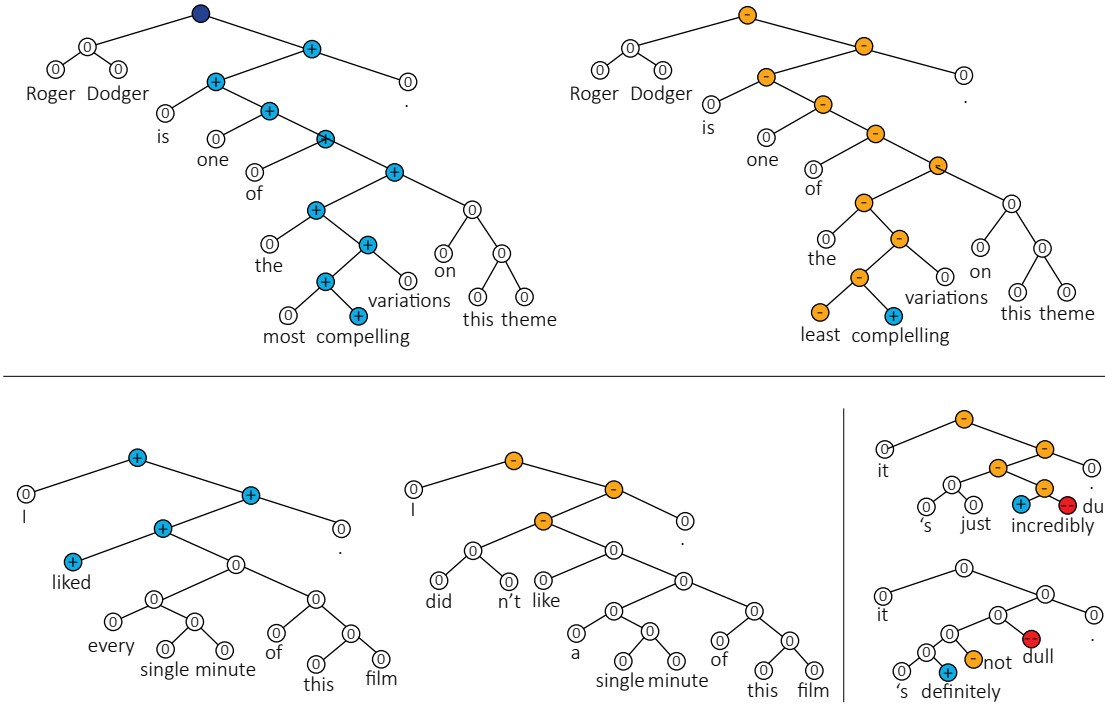
Kolejny zbiór stworzony na Uniwersytecie Stanforda. Skupia się na zagadnieniu analizy wydźwięku złożonych zdań, adresując istotny problem – wydźwięk pojedynczych słów nie zawsze koresponduje z wydźwiękiem dłuższej wypowiedzi. Oparty został na bazie recenzji filmów, zamieszczonych przez użytkowników w serwisie Rotten Tomatoes[70]. Badacze podzielili zebrane recenzje na frazy, które zostały później sklasyfikowane przez ludzi z wykorzystaniem 25-stopniowej skali (pozwalającej na przyznanie oceny od „bardzo negatywne" przez „neutralne" do „bardzo pozytywne"). Repozytorium liczy ponad 215 tysięcy opisanych fraz.

- **Sentiment140**[71]

Dane udostępnione przez twórców usługi o tej samej nazwie pozwalającej na przeglądanie wyników analizy wydźwięku opublikowanych na Twitterze wiadomości, dotyczących konkretnego tematu, marki lub produktu. Zawierają

67 http://ai.stanford.edu/~amaas/data/sentiment/
68 https://www.imdb.com/
69 https://nlp.stanford.edu/sentiment/code.html
70 https://www.rottentomatoes.com/
71 http://help.sentiment140.com/

RYSUNEK 4.15. **PRZYKŁAD ANALIZY WYDŹWIĘKU ZDAŃ PODZIELONYCH NA POJEDYNCZE SŁOWA**



Źródło: [162]

treść wiadomości, nick autora, datę oraz ocenę wydźwięku w trzystopniowej skali (negatywne, neutralne, pozytywne). W przeciwieństwie do poprzednich zbiorów, w tym przypadku ocena wydźwięku została dokonana automatycznie, bez udziału ludzi, w oparciu o słowa kluczowe oraz emotikony.

4.5.3. Przetwarzanie języka naturalnego

- **Google Books n-grams**[72]

Opublikowany przez Amazon zbiór n-gramów wygenerowanych na podstawie zbiorów Google Books[73]. Zawiera ponad 2 terabajty danych, przygotowanych do przetwarzania z wykorzystaniem technologii big data (Hadoop). N-gram to model analizy języka, umożliwiający przewi-

72 https://aws.amazon.com/datasets/googlebooks-ngrams/
73 https://books.google.com/

dywanie wystąpień określonej sekwencji znaków w tekście. Przy wykorzystaniu odpowiednio dużego zbioru danych wejściowych, n-gramy pozwalają na stworzenie modeli o bardzo dużych możliwościach predykcyjnych. Wykorzystywane są nie tylko do analizy tekstu, ale również rozpoznawania mowy. Google Books n-grams zawiera 11 podzbiorów, podzielonych według języków (w tym kilka podzbiorów dla języka angielskiego np. angielski brytyjski i amerykański).

- **Amazon reviews**[74]

Zbiór recenzji produktów zamieszczonych przez użytkowników w serwisie Amazon[75]. Zawiera ponad 142,8 miliona recenzji, napisanych w latach 1996-2014. Poza tekstem recenzji, zamieszone zostały również informacje o autorze oraz przyznana produktowi ocena i metadane produktu (np. cena, marka). Ze względu na swoją charakterystykę, zbiór ten ma liczne zastosowania takie jak analiza wydźwięku oraz rozpoznawanie charakterystycznych cech języka konkretnych użytkowników.

- **Yelp reviews**[76]

Zestaw ponad 6,6 miliona recenzji pochodzących z serwisu Yelp[77]. Wyróżnia go fakt, iż nie dotyczy konkretnych produktów a usług. Recenzje, napisane przez 1,6 miliona użytkowników, opisują niemal 200 tysięcy usługodawców. Zbiór zawiera również 200 tysięcy zdjęć. Dane mogą zostać wykorzystane do analizy tekstu, wydźwięku, a także przetwarzania obrazów (np. klasyfikacja rodzajów jedzenia lub ocena estetyki). Dane udostępnione zostały na potrzeby edukacyjne i badawcze. Związany jest z nimi konkurs, opisany w podrozdziale **5.2.6**.

Krótkie podsumowanie najważniejszych informacji o przedstawionych zbiorach danych znajduje się w tabeli **4.3**.

74 http://jmcauley.ucsd.edu/data/amazon/
75 https://www.amazon.com/
76 https://www.yelp.com/dataset
77 https://www.yelp.com

4.5.4. Podsumowanie

Choć algorytmy uczenia maszynowego do poprawnego działania wymagają danych wysokiej jakości, nieprawdą jest, iż zawsze niezbędne są zbiory dużej wielkości. Dzięki zastosowaniu techniki *transfer learningu* [**138**] możliwe jest stworzenie modeli o dużych zdolnościach predykcyjnych w oparciu o niewielkie wolumeny danych. Transfer learning pozwala na wykorzystanie modeli wcześniej wytrenowanych i dostosowanie ich do swoich potrzeb podczas dodatkowej, krótszej fazy uczenia. W praktyce istnieją dwie metody wykorzystania tej techniki.

Można skorzystać z dostępnych publicznie zasobów danych, takie jak opisane powyżej, w celu wstępnego przygotowania modeli, po czym dostosować je do specyficznych zastosowań z wykorzystaniem własnych danych. Proces ten pozwala na całkowitą kontrolę nad tworzoną architekturą, wymaga jednak odpowiedniej infrastruktury sprzętowej (np. kart graficznych, na których obliczenia uczenia maszynowego przeprowadzane są najwydajniej) oraz czasu (trening rozbudowanych architektur na największych zbiorach może trwać nawet kilka tygodni).

Drugim wyjściem jest skorzystanie z gotowych wag (parametrów obliczanych w procesie uczenia modelu). Gotowe do implementacji modele udostępniane są często przez badaczy. Repozytorium biblioteki **TensorFlow**[78] zawiera wiele implementacji publikowanych przez naukowców architektur głębokich sieci neuronowych. Serwis **ModelDepot**[79] specjalizuje się w oferowaniu rozwiązań uczenia maszynowego gotowych do wdrożenia. Z kolei twórcy biblioteki **fastai**[80] [i] dali możliwość tworzenia wytrenowanych modeli w ramach API samej biblioteki.

Zasoby te sprawiają, że technologie uczenia maszynowego stają się coraz bardziej przystępne dla przedsiębiorców, którzy nie mogą pozwolić sobie na zatrudnienie specjalistów w tej dziedzinie. Pozwalają także na redukcję kosztów związanych z opracowaniem architektury, zbieraniem danych oraz zakupem niezbędnej infrastruktury.

Należy jednak pamiętać, iż zbiory danych, architektury i wagi, publikowane w Internecie, mogą być objęte licencjami, które dokładnie określają dopuszczalne zastosowania. Złamanie tych zasad równoznaczne jest z naruszeniem praw autorskich, dlatego w interesie przedsiębiorców i innych potencjalnych użytkowników leży kontrolowanie wykorzystywanych zasobów również pod tym względem.

78 https://github.com/tensorflow/models/
79 https://modeldepot.io/
80 https://github.com/fastai/fasta

TABELA 4.3. **ZBIORY DANYCH DO GŁĘBOKIEGO UCZENIA MASZYNOWEGO – PODSUMOWANIE**

NAZWA	AUTORZY	PRZEZNACZENIE	ROZMIAR	LICENCJA /ZASTOSOWANIA
LabelMe	Massachusetts Institute of Technology	Wykrywanieisegmentacja obiektów	178 371 zdjęć	Licencja MIT.
ImageNet	Uniwersytet Stanforda, Uniwersytet w Princeton	Rozpoznawanie i lokalizacja obiektów	14 197 122 zdjęcia	Nieokreślone. Prawa do zdjęć posiadają ich autorzy
COCO- Common Objects in Context	Microsoft	Wieloaspektowa analiza obrazów	330 000 zdjęć	Opisy do zdjęć na licencji Creative Commons Attribution 4.0. Prawa do zdjęć posiadają ich autorzy.
COIL-100	Uniwersytet Columbia	Rozpoznawanie obiektów	7200 zdjęć	Nieokreślone.
Labeled Faces in the Wild	Uniwersytet Massachusetts	Rozpoznawanie twarzy	13 233 zdjęcia	Nieokreślone.
Open Images	Google	Rozpoznawanie i lokalizacja obiektów	ok. 9 000 000 zdjęć	Opisy na licencji Creative Commons Attribution 4.0. Zdjęcia na licencji Creative Commons Attribution 2.0.
Indoor Scene Recognition	Massachusetts Institute of Technology	Klasyfikacja wnętrz budynków	15 620 zdjęć	Do zastosowań badawczych.
Flickr 8K	Uniwersytet Illinois	Generowanie opisów scen	8092 zdjęcia	Tylko dla zastosowań badawczych/ edukacyjnych.
IMDB Reviews	Uniwersytet Stan-forda	Binarna klasyfikacja wydźwięku tekstu	150 000 recenzji	Nieokreślone.
Stanford Sentiment Treebank	Uniwersytet Stan-forda	Klasyfikacja wydźwięku tekstu	215 154 frazy	Nieokreślone.
Sentiment140	Sentiment140	3-stopniowa analiza wydźwięku tekstu	1 600 000 tweetów	Do zastosowań badawczych.
Google Books ngrams	Amazon	Analiza języka naturalnego	2TB danych tekstowych	Licencja Creative Commons Attribution 3.0.
Amazon reviews	Uniwersytet Stan-forda	Analiza języka naturalnego	142,8 miliona recenzji	Nieokreślone.
Yelp reviews	Yelp	Analiza języka naturalnego i zdjęć	6 685 900 recenzji, 200 000 zdjęć	Tylko dla zastosowań ba-dawczych/edukacyjnych.

Źródło: Opracowanie własne.

4.6. Analiza przypadków

4.6.1. OpenStreetMap

Dane z OpenStreetMap (dalej PSM) dostępne na licencji Open Database License[81] (ODbL), która m.in. oznacza, że jeżeli chcemy publikować prace oparte na tych danych, należy wskazać OSM jako źródło, a jeżeli zmieniamy (poprawiamy) te dane, musimy udostępnić te zmienione dane do społeczności [**134**]. Na przykład potrzeba w dokonaniu modyfikacji tych otwartych danych geograficznych może się pojawić w przypadku, gdy:

- Chcemy zmodyfikować wygląd mapy, aby pasowała do naszej własnej marki.
- Nie chcemy wyświetlać reklam biznesowych dostępnych na mapach znanych wyszukiwarek.
- Spodziewamy się, że wolumen użycia będzie zbyt duży, aby mógł być obsługiwany przez serwery OSM.
- Chcemy wyróżnić określone typy funkcji lub usunąć nieprecyzyjne szczegóły, nie renderując niektórych typów funkcji.
- Mini-mapa przedstawiająca lokalizację twojej firmy względem jednej lub dwóch lokalnych ulic.
- Mapa w tle z czymś niegeograficznym, na przykład logo.
- Grafiki lub mapy na podstawie danych geodezyjnych OSM.

Dane z OSM dają wiele możliwości do wzbogacenia oraz usprawnienia istniejących aplikacji. Przykłady wdrożenia OSM [**135**]:

- http://opencyclemap.org/ – pokazywanie informacji przydatnych rowerzystom (np. ścieżki rowerowe, parkingi, sklepy).
- https://wheelmap.org/ – pozwala znaleźć miejsca dostępne dla wózków inwalidzkich na całym świecie.
- https://www.komoot.de/ – aplikacja pod urządzenia z systemem operacyjnym Android do nawigacji dla rekreacji i sportów na świeżym powietrzu.
- https://www.openstreetbrowser.org/ – aplikacja do przeglądania informacji w wyświetlonej części mapy. Celem projektu jest dostarczenie wysoce dynamicznej mapy, która sprawia, że każda zmapowana funkcja jest łatwo dostępna dla użytkownika.
- http://öpnvkarte.de – wyświetla na światowej mapie obiekty transportu publicznego, dzięki czemu nie ma konieczności przeglądania witryn poszczególnych operatorów.
- http://openseamap.org – serwis zbiera dane żeglarskie oraz geoprzestrzenne w celu utworzenia mapy morskiej na całym świecie.
- https://www.wanderreitkarte.de/ – mapa jazdy konnej i turystyki pieszej.
- https://hiking.waymarkedtrails.org/ – mapy dla turystyki pieszej.

81 https://opendatacommons.org/licenses/odbl/index.html

- http://www.openfiremap.org/ – mapy o pożarach, która zawiera m.in. stacje przeciwpożarowe, hydranty, zbiorniki wodne i stawy wykorzystywane do gaszenia pożarów.
- https://www.xctrails.org – pokazuje trasy narciarskie oraz trasy biegowe. Zawiera również informacje dodatkowe z innych źródeł, takich jak pogoda, stoki, profile wysokości itp.
- https://www.4umaps.com – mapy topograficzne zewnętrznie, wędrówki oraz mapy rowerowe. Może być przydatne do pieszych wycieczek, kolarstwa górskiego, jazdy na rowerze, wspinaczki itp. Mapy zawierają linie elewacji, cieniowanie gór, wysokość i nazwę szczytu, ulice, drogi, ścieżki i szlaki, a także źródła, supermarkety, restauracje, hotele, schroniska itp.
- https://www.openrailwaymap.org/ – to szczegółowa mapa online światowej infrastruktury kolejowej.

4.6.2. Babelnet

BabelNet to wielojęzyczny słownik encyklopedyczny, zawierający leksykograficzne i encyklopedyczne opisy pojęć. Łączy pojęcia i nazwy podmiotów w bardzo dużą sieć relacji semantycznych, za pomocą wpisów zwanych synsetami [**127**]. Każdy synset reprezentuje jakieś znaczenie i zawiera wszystkie synonimy wyrażające to znaczenie w różnych językach.

BabelNet 4.0 zawiera niemal 16 milionów synsetów (w tym ponad 3,5 miliona w języku polskim), opisujących ponad 9,6 miliona obiektów. Stworzona przez nie sieć semantyczno-leksykalna składa się z 277 milionów relacji. Koncepty ilustrowane są zdjęciami, których obecnie jest ponad 54 miliony. Babelnet obejmuje ponad 280 języków i jest automatycznie wzbogacany za pośrednictwem takich otwartych baz wiedzy jak Wikipedia, WordNet, OmegaWiki, Wikisłownik, Wikidane, Wikicytaty, verbNet, GeoNames, ImageNet, czy WN-MAP. Dane pochodzą łącznie z 47 różnych źródeł.

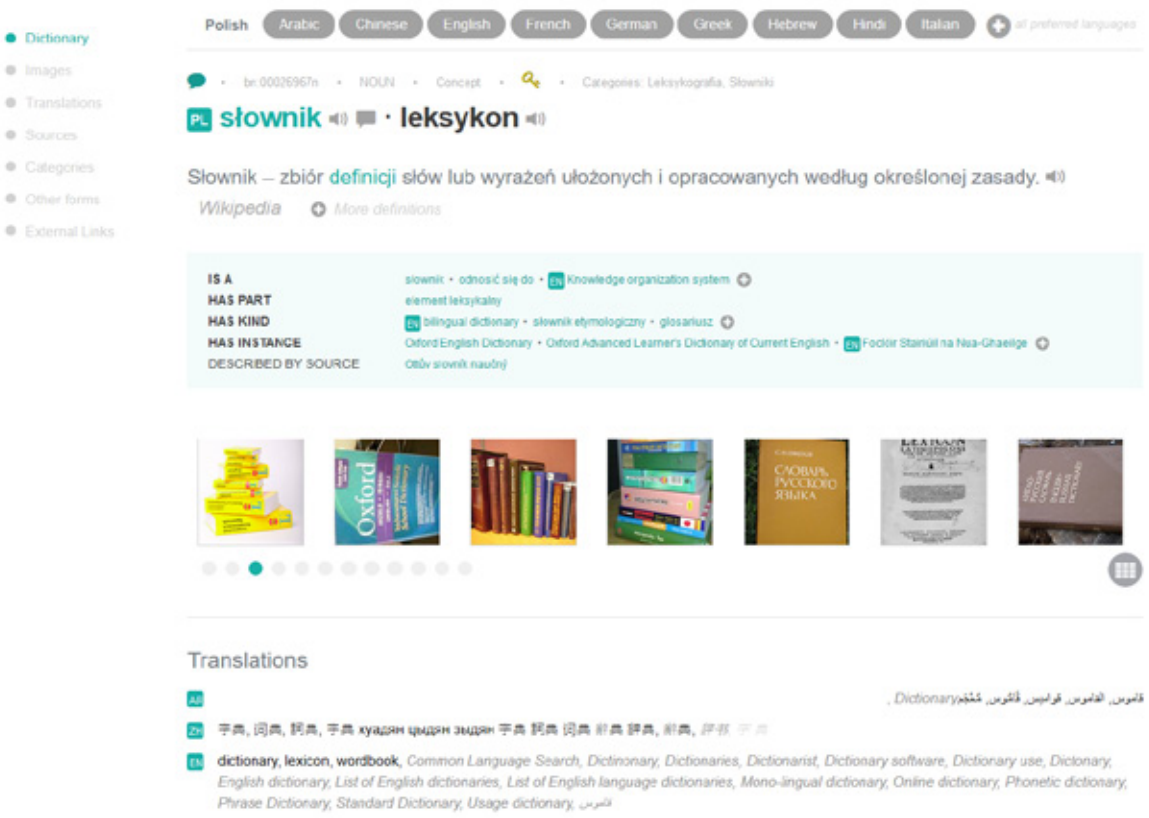
Ponadto Babelnet zintegrowany jest z usługą Babelfly[82], pozwalającą na semantyczną analizę tekstu w oparciu o grafy, oraz MultiWiBi[83] [**53**], taksonomią Wikipedii. Istnieje również możliwość skorzystania z usługi za pośrednictwem kilku wersji API (HTTP, Java, Python, SPARQL). Dostęp jest darmowy, ale ograniczony limitem zapytań, który może zostać nieodpłatnie usunięty na potrzeby projektów badawczych. Przykładowy wynik wyszukiwania w Babelnet przedstawiony został na rys. **4.16**.

Babelnet dostarcza narzędzi pomocnych przy próbach rozwiązania wielu istotnych problemów z zakresu przetwarzania języka naturalnego. Najwięcej dotychczas przeprowadzonych badań dotyczyło ujednoznaczniania znaczenia słów [**124**, **128**, **129**]. Zagadnienie to bada możliwość określenia poprawnego znaczenia wieloznacznego słowa w konkretnym kontekście.

82 http://babelfy.org/
83 http://www.wibitaxonomy.org/

Ma ono ogromne znaczenie dla usprawnienia działania wyszukiwarek internetowych, botów komunikujących się z ludźmi czy narzędzi do automatycznej analizy tekstu.

RYSUNEK 4.16. **PRZYKŁADOWY WYNIK WYSZUKIWANIA DLA HASŁA *SŁOWNIK***



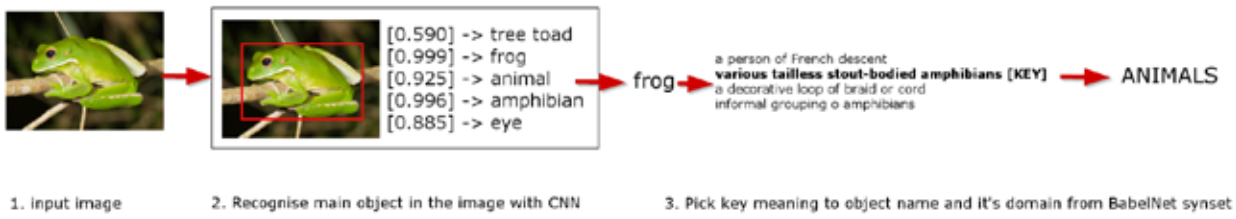
Źródło: https://babelnet.org

Przykład praktycznego – z punktu widzenia biznesu – zastosowania możliwości Babelnetu zaprezentowali Wieczorek, Filipiak i Filipowska [**191**]. Autorzy połączyli model rozpoznawania obiektów na zdjęciach (stworzony za pomocą technik głębokiego uczenia) i Babelnet, w celu stworzenia prototypu systemu profilującego potencjalne zainteresowania użytkownika na podstawie aktywności w mediach społecznościowych. System ten, za pomocą głębokiej sieci neuronowej, rozpoznawał obiekty zawarte na zdjęciach, z którymi użytkownik wszedł w interakcję w serwisie Flickr[84]. Uzyskane w ten sposób predykcje przekazywane były do Babel-

84 https://www.flickr.com/

netu za pomocą API dla języka Python, w celu uzyskania bardziej ogólnych kategorii. Kategorie te wybierane były za pomocą systemu znaczeń kluczowych, zaimplementowanego w sieć semantyczną Babelnet. Zabieg ten miał na celu uniknięcie problemów wynikających z bardzo wysokiej szczegółowości predykcji uzyskiwanych przez modele trenowane na zbiorze danych ImageNet (patrz podrozdział **4.5.1**). Dzięki temu możliwe było wykorzystanie dostępnych publicznie wcześniej wytrenowanych modeli, bez przeprowadzania dodatkowego treningu na własnym zbiorze danych. Implementacja komunikacji z Babelnet za pomocą API jest zdecydowanie mniej problematyczna niż przygotowywanie dedykowanego modelu sieci neuronowej. Przykład ten pokazuje, iż łączenie ze sobą publicznie dostępnych narzędzi może pozwolić na obniżenie bariery dostępu do niektórych rozwiązań dla osób, które nie są specjalistami w dziedzinie uczenia maszynowego. Profilowanie zainteresowań może zaś znaleźć zastosowanie w wielu dziedzinach, m.in. w badaniach rynku i marketingu.

RYSUNEK 4.17. **PRZEBIEG PROCESU GENERALIZACJI NAZW OBIEKTÓW Z WYKORZYSTANIEM SIECI NEURONOWEJ I BABELNET**



Źródło: [191]

4.6.3. Common Crawl

Duże ilości danych, dostępne w ramach Common Crawl, mogą być pomocne przy rozwiązywaniu problemów definiowanych w wielu obszarach przez przedsiębiorstwa. Do najistotniejszych można zaliczyć:

- Analiza sieci. Zrozumienie powiązań między stronami WWW, potencjalnie wskazującymi na połączenia np. między ludźmi czy firmami.
- Zbudowania i przetestowanie i własnej wyszukiwarki internetowej na gotowych danych, dedykowanej określonej dziedzinie.
- Przetwarzanie dokumentów. W niektórych obszarach informacje nie zmieniają się zbyt często, na przykład: strony rządowe, dokumenty, ustawy. Te dane można użyć do zbudowania usługi bez konieczności budowania własnego robota internetowego.

Dobrym przykładem wyszukiwarki oferującej swobodny dostępny interfejs wyszukiwania dla dwóch korpusów ClueWeb oraz Common Crawl jest ChatNoir [**21**]. W przeciwieństwie do

większości komercyjnych wyszukiwarek, oferuje również interfejs API, który jest dostępny bezpłatnie dla naukowców. Kod źródłowy wyszukiwarki jest dostępny na GitHub[85].

Innym przykładem wykorzystania danych z Common Crawl jest analiza rozpowszechnienia reklam internetowych [**32**]. Wyniki pokazują, że około 53% z serwisów wyświetla reklamy. Przy tym, można otrzymać szczegółowe informację dla każdej z blisko 23 mln domen na temat wykorzystywanych reklam. Np. dla polskich domen (z rozszerzeniem „.pl") reklamy wyświetlają 167 tyś. witryn, natomiast nie zidentyfikowano żadnych reklam na ok. 133 tyś. polskich serwisów internetowych. Program do szukania reklam jest napisany w języku Java oraz dostępny jest jego kod źródłowy[86].

Wyszukiwanie serwisów różnego typu

CommonCrawl może być pomocny do identyfikacji serwisów internetowych określonego typu. To może być pomocne w celu agregacji danych z takich witryn, ponieważ udostępniają one podobny zakres narzędzi i usługi (w tym interfejsy API). Na przykład, jeżeli szukamy serwisów typu wiki, możemy skorzystać z katalogów, które mają listę takich serwisów[87]. Podstawową wadą tych katalogów jest to, że zazwyczaj one są wzbogacane ręcznie, i mogą zawierać nieaktualnie dane (niektóre serwisy już nie działają, a z listy one nie zostały usunięte). Żeby ten proces zautomatyzować, musimy mieć chociażby jedną stronę z serwisu wiki, żeby zidentyfikować oprogramowanie, które zostało wykorzystane w tym serwisie do generowania tej strony. Najczęściej serwisy wiki wykorzystują wolne oprogramowanie MediaWiki[88], które domyślnie dodaje oznaczenie w meta-danych strony o nazwie oprogramowania oraz jego wersji. Na przykład, w momencie powstania niniejszego raportu na dowolnej stronie Wikipedii można było znaleźć w kodzie HTML wskazany niżej wiersz:

```
1   <meta name="generator" content="MediaWiki 1.33.0-wmf.14"/>
```

Meta dane przedstawione w notacji HTML przechowywane razem z kodem źródłowym strony w formacie WARC w CommonCrawl. Ten kod zostaje przekształcony do formatu JSON w plikach WAT. W przypadku meta tagu „generator" na stronach Wikipedii, będzie on miał kod podany poniżej:

```
1   {"generator":"MediaWiki 1.33.0-wmf.14"}
```

85 https://github.com/chatnoir-eu
86 https://bitbucket.org/yymoto/common-crawl-archivead-analyser/src/master/
87 https://www.mediawiki.org/wiki/Sites_using_MediaWiki/en
88 https://www.mediawiki.org/wiki/MediaWiki

Mając dane w formacie WAT wygenerowane przez CommonCrawl, możemy identyfikować serwisy po metatagu „generator", który jest przypisywany stronom wygenerowanym przy pomocy MediaWiki lub też innego oprogramowania do tworzenia serwisów typu wiki. Dane w formacie WAT zajmują około 19 TB w postaci skompresowanej[89]. Nie ma jednak konieczności pobierania wszystkich tych danych jednocześnie – można je przetwarzać w pakietach. Dane te są podzielone na około 64 tysiące plików, co daje możliwość pracy przyrostowo na stosunkowo niewielkich zbiorach danych.

Przed rozpoczęciem analizy należy wybrać zbiór danych na odpowiednią datę. Pełna lista zbiorów Common Crawl dla różnych lat i miesięcy dostępna na stronie: http://commoncrawl.org/the-data/get-started/. Na przykład, chcemy wybrać zbiór danych wygenerowany w styczniu 2019 roku, wtedy możemy przejść do odpowiedniej strony: https://commoncrawl.s3.amazonaws.com/crawldata/CC-MAIN-2019-04/index.html. Na tej stronie musimy wybrać listę plików z danymi w formacie, którym jesteśmy zainteresowani – w naszym przykładzie to jest format WAT i listę plików w tym formacie można pobrać tu: https://commoncrawl.s3.amazonaws.com/crawl-data/CC-MAIN-9-04/wat.paths.gz. W tym pliku w każdej linijce podana jest ścieżka do każdego z 64 tysięcy plików, w którym zawarte są meta informacje w formacie JSON. Wycinek z tej listy pokazany na **4.3**.

KOD ŹRÓDŁOWY 4.3. **WYCINEK Z LISTY PLIKÓW W FORMACIE WAT ZGENEROWANY PRZEZ COMMON CRAWL W STYCZNIU 2019 ROKU**

```
1  ...
2  crawl-data/CC-MAIN-2019-04/segments/1547583656...38-00002.warc.wat.gz
3  crawl-data/CC-MAIN-2019-04/segments/1547583656...38-00003.warc.wat.gz
4  crawl-data/CC-MAIN-2019-04/segments/1547583656...38-00004.warc.wat.gz ...
```

W celu uzyskania pełnej ścieżku do każdego z wymienionych na liście plików należy dodać na początku protokół oraz domenę „https://commoncrawl.s3.amazonaws.com/". Każdy z tych plików na liście zajmuje około 200-300 MB i może być analizowany niezależnie od innych. Przykład programu, który przyrostowo pobiera pliki w formacie WAT, ekstrahuje niezbędne informacje zawarte w metatagu „generator" oraz zapisuje wyniki dla każdego z analizowanych plików, co umożliwia kontynuacje pracy nawet po przerywaniu procesu, można znaleźć na stronie https://github.com/Lewoniewski/CommonCrawl.

89 dane ze stycznia 2019 roku

4.6.4. DBpedia

Materiały z Wikipedii są stosunkowo często wykorzystywane na innych stronach internetowych, czy to przez cytowanie czy bezpośrednie użycie tekstu. Istnieją metody pozwalające na wykrywanie struktur tekstowych w sieci WWW, które są podobne do treści zawartych w artykułach Wikipedii [**6**].

Istnieją różne narzędzia, pozwalające na ekstrakcję treści artykułów tej encyklopedii. Jednym z przykładów takich narzędzi jest WikiExtractor[90], które to generuje zwykły tekst ze zrzutu bazy danych Wikipedii, odrzucając wszelkie inne informacje lub adnotacje znajdujące się na stronach Wikipedii, takie jak obrazy, tabele, odnośniki oraz listy. Każdy dokument w kopii zapasowej tej encyklopedii jest reprezentowany jako pojedynczy element XML, przedstawiony w listingu **4.4.**

KOD ŹRÓDŁOWY 4.4. **WYCINEK Z KODU ŹRÓDŁOWEGO W FORMACIE XML ARTYKUŁU POLSKOJĘZYCZNEJ WIKIPEDII Z KOPII ZAPASOWEJ**

```
<page>
    <title>Polska</title>
    <ns>0</ns>
    <id>44895</id>
    <revision>
        <id>55830085</id>
        <parentid>55829975</parentid>
        <timestamp>2019-02-06T22:09:17Z</timestamp>
        <contributor>
            <username>Urabura</username>
            <id>932433</id>
        </contributor>
        <comment>Anulowanie wersji 55829975 ....</comment>
        <model>wikitext</model>
        <format>text/x-wiki</format>
        <text xml:space="preserve" bytes="267096">{{Inne znaczenia|państwa}}
            {{Państwo infobox
            |nazwa_oryginalna = Rzeczpospolita Polska
            |nazwa_polska =
            |flaga_obraz = Flag of Poland.svg
            |godło_obraz = Herb Polski.svg
            |nazwa_dopełniacz = Polski
            |p1 = Polska Rzeczpospolita Ludowa
            |p1_flaga = Flag of Poland (1928-1980).svg
            ...
            [[Kategoria:Polska| ]]
            [[Kategoria:Członkowie Organizacji Narodów Zjednoczonych]]
            [[Kategoria:Państwa członkowskie Unii Europejskiej]]
            [[Kategoria:Państwa należące do NATO]]
            [[Kategoria:Hasła kanonu polskiej Wikipedii]]
        </text>
        <sha1>28noqhw9gk1skotq7org9woe1ee9t7y</sha1>
    </revision>
</page>
```

90 https://github.com/attardi/wikiextractor

Przed rozpoczęciem pracy z programem WikiExtractor należy pobrać kopię zapasową wybranej wersji językowej Wikipedii. Na przykład dla polskojęzycznej Wikipedii ostatnia wersja pliku ze wszystkimi artykułami w wersji XML znajduje się pod adresem: https://dumps.wikimedia.org/plwiki/latest/plwikilatest-pages-articles.xml.bz2.

W celu uruchomienia skryptu należy podać nazwę pliku oraz dodatkowe opcje, tj. WikiExtractor.py [opcje] plik-kopii-zapasowej-xml. Po jego wykonaniu otrzymamy dane wyjściowe w postaci wielu plików o podobnej wielkości, zapisanych w bieżącym katalogu. Każdy plik będzie zawierał kilka dokumentów w formacie pokazanym w listingu **4.5**.

KOD ŹRÓDŁOWY 4.5. **WYCINEK Z KODU OTRZYMANEGO PRZY POMOCY WIKIEXTRACTOR Z KOPII ZAPASOWEJ POLSKOJĘZYCZNEJ WIKIPEDII**

```
1  <doc id="44895" revid="55830085" url="https://pl.wikipedia.org/wiki/Polska" title="Polska">
2      Polska, Rzeczpospolita Polska (RP) - państwo unitarne w Europie
3      Środkowej, położone między Morzem Bałtyckim na północy a Sudetami
4      i Karpatami na południu, w przeważającej części w dorzeczu Wisły
5      i Odry. Od północy Polska graniczy z...
6  </doc>
```

Element „doc" w otrzymanym kodzie XML posiada następujące atrybuty:

- id – identyfikuje dokument za pomocą unikalnego numeru seryjnego
- revid – unikalny numer edycji dokumentu.
- url – podaje adres URL artykułu Wikipedii.
- title – nazwa artykułu Wikipedii.

Na potrzeby uniwersalnej ekstrakcji danych ustrukturyzowanych i nieustrukturyzowanych z Wikipedii jednym z najlepszych narzędzi jest DBpedia Ekstraction Framework[91], które można uruchomić na własnym komputerze. To narzędzie pozwala na ekstrakcje semantycznych trójek z kopii zapasowych Wikipedii[92]. Wynik działania ekstraktorów DBpedii na podstawie kopii zapasowych Wikipedii na koreślone daty można pobrać na stronie https://wiki.dbpedia.org/Datasets.

Można korzystać również z niektórych ekstraktorów DBpedii za pośrednictwem udostępnionych usług sieciowych. Na przykład w celu ekstrakcji strukturyzowanych danych określonego artykułu Wikipedii można wykorzystać zapytanie URL wg schematu pokazanego w listingu **4.6**.

91 https://github.com/dbpedia/extraction-framework
92 https://dumps.wikimedia.org/backup-index.html

KOD ŹRÓDŁOWY 4.6. **SCHEMAT ZAPYTANIA URL DO EKSTRAKCJI DANYCH Z ARTYKUŁU WIKIPEDII W POSTACI SEMANTYCZNYCH TRÓJEK DBPEDII**

```
1  http://mappings.dbpedia.org/server/extraction/{Kod wersji
2      ↳ językowej}/extract?title={Nazwa Artykułu
3      ↳ Wikipedii}&revid=&format=turtle-triples
```

Otrzymane dane zawierają notację DBpedii, która jest zależna od reguł mapowania. Na przykład dla zapytania artykułu o nazwie „Warsaw" w angielskiej Wikipedii otrzymamy wynik częściowo pokazany w listingu **4.7**.

KOD ŹRÓDŁOWY 4.7. **WYNIK DZIAŁANIA EKSTRAKTORA DBPEDII WYWOŁANY PRZEZ URL ZGODNIE Z LISTINGIEM 4.6**

```
1  # started 2019-02-12T06:01:41Z
2  <http://en.dbpedia.org/resource/Warsaw> <http://www.w3.org/2000/01/rdf-schema#label>
3      ↳ "Warsaw"@en .
4  <http://en.dbpedia.org/resource/Warsaw>
5      ↳ <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
6      ↳ <http://dbpedia.org/ontology/City> .
7  <http://en.dbpedia.org/resource/Warsaw>
8      ↳ <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://schema.org/City>  .
9  <http://en.dbpedia.org/resource/Warsaw>
10     ↳ <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ,
11 <http://www.wikidata.org/entity/Q515> .
12 <http://en.dbpedia.org/resource/Warsaw>
13     ↳ <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ,
14     ↳ <http://dbpedia.org/ontology/Settlement> .
15 <http://en.dbpedia.org/resource/Warsaw>
16     ↳ <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ,
17     ↳ <http://www.wikidata.org/entity/Q486972> .
18 <http://en.dbpedia.org/resource/Warsaw>
19     ↳ <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ,
20     ↳ <http://dbpedia.org/ontology/PopulatedPlace> .
21 <http://en.dbpedia.org/resource/Warsaw>
22     ↳ <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ,
23     ↳ <http://dbpedia.org/ontology/Place> .
24 <http://en.dbpedia.org/resource/Warsaw>
25     ↳ <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://schema.org/Place> .
26 . . .
27 # completed 2019-02-12T06:01:42Z
```

4.6.5. Wikidane

Podobnie jak Wikipedia, w Wikidanych przechowywane są różne typy danych (np. tekst, obrazy, współrzędne, mapy, daty). Aby uzyskać dane z tej bazy wiedzy, wystarczy użyć trójek w zapytaniu SPARQL. Zobaczmy, jak może wyglądać takie zapytanie. Na przykład, jeżeli chcemy znaleźć wszystkie miasta Polski opisane w Wikidanyh, to szukany obiekt musi spełniać co najmniej dwa warunki, które mają w Wikidanych odpowiednie nazwane cechy (w nawiasach podane te nazwy):

- „jest to" (P31) – „miasto" (Q515),
- „państwo" (P17) – „Polska" (Q36).

Innymi słowy w zapytaniu do Wikidanych używamy określonych identyfikatorów, aby zdefiniować właściwą relację oraz element. W języku SPARQL powyższe zapytanie zapiszemy w postaci przedstawionej w listingu **4.8**.

KOD ŹRÓDŁOWY 4.8. **ZAPYTANIE SPARQL, KTÓRE WYSZUKUJE WSZYSTKIE MIASTA POLSKI OPISANE W WIKIDANYCH**

```
SELECT ?miasto WHERE {
    ?miasto wdt:P31 wd:Q515.
    ?miasto wdt:P17 wd:Q36.
}
```

W wyniku uruchomienia tego kodu w Wikidata Query Service[93] dostaniemy listę ponad 900 identyfikatorów elementów (np. Q268 to jest „Poznań"), które pasują do zapytania. Te identyfikatory nie są jednak czytelne dla człowieka. Dlatego można wprowadzić dodatkowe kolumny, które pokażą również etykiety (nazwy) w języku polskim do poszczególnych elementów. Dodatkowo uwzględnimy populacje każdego miasta – ograniczymy ich listę t, w których mieszka co najmniej 100 tys. osób oraz dodamy kolumnę, gdzie będzie pokazana ta wartość.

KOD ŹRÓDŁOWY 4.9. **ZAPYTANIE SPARQL WYSZUKUJĄCE WSZYSTKIE MIASTA POLSKI W WIKIDANYCH, POSIADAJĄCE PONAD 100 TYS. MIESZKAŃCÓW, Z SORTOWANIEM WEDŁUG LICZBY MIESZCZKANCÓW**

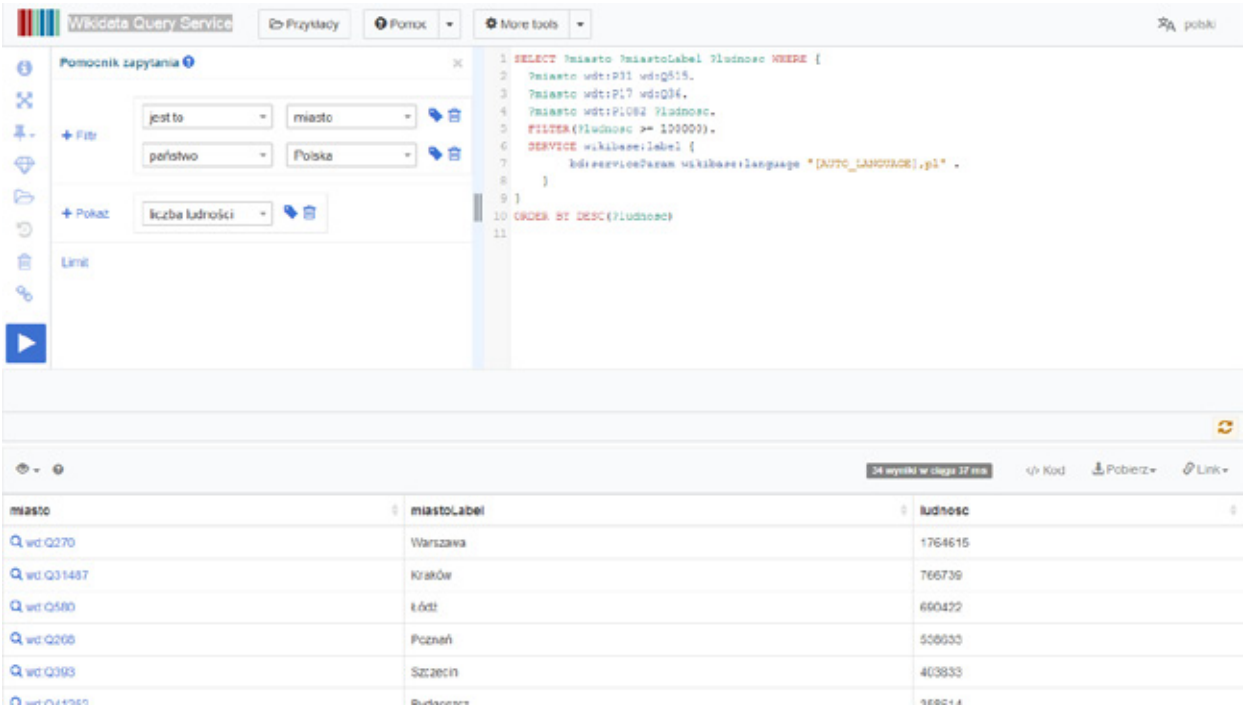
```
SELECT ?miasto ?miastoLabel ?ludnosc WHERE {
    ?miasto wdt:P31 wd:Q515.
    ?miasto wdt:P17 wd:Q36.
    ?miasto wdt:P1082 ?ludnosc.
    FILTER(?ludnosc >= 100000).
    SERVICE wikibase:label {
        bd:serviceParam wikibase:language "[AUTO_LANGUAGE],pl" .
    }
}
ORDER BY DESC(?ludnosc)
```

93 https://query.wikidata.org/

Warto zaznaczyć, że identyfikatory cech oraz elementów są opisane za pośrednictwem pomocnika zapytania, dostępnego w ramach Wikidata Query Service. Interfejs tego narzędzia z przedstawionym zapytaniem wraz z wynikami pokazano na rys. **4.18**.

Wyniki wyszukiwania mogą być przedstawione inaczej niż w wersji tabelarycznej. Na przykład możemy pokazać wybrane miasta na mapie. Wprowadzamy zmiany poprzez nieznaczące zmiany w zapytaniu – nowy kod zapytania przedstawiony na **4.10**. Nowymi w stosunku do poprzedniego kodu są wierze 1 oraz 8. Dodatkowo została dopisana zmienna „?geo" w wierszu 2. Otrzymana mapa jest pokazana na rys. **4.19** oraz została zbudowana w oparciu o dane z OpenStreetMap (por. **4.3.7**).

RYSUNEK 4.18. **WIKIDATA QUERY SERVICE Z WYNIKAMI ZAPYTANIA 4.9**



Źródło: https://query.wikidata.org

KOD ŹRÓDŁOWY 4.10. **ZAPYTANIE SPARQL, KTÓRE UMOŻLIWIA POKAZANIE NA MAPIE POLSKICH MIAST POSIADAJĄCYCH PONAD 100 TYS. MIESZKAŃCÓW**

```
#defaultView:Map
SELECT ?miasto ?miastoLabel ?ludnosc ?geo WHERE {
    ?miasto wdt:P31 wd:Q515.
    ?miasto wdt:P17 wd:Q36.
    ?miasto wdt:P1082 ?ludnosc.
    SERVICE wikibase:label {
        bd:serviceParam wikibase:language "[AUTO_LANGUAGE],pl" .
    }
    FILTER (?ludnosc >= 100000).
    OPTIONAL { ?miasto wdt:P625 ?geo. }
}
ORDER BY DESC (?ludnosc)
```

Widok wyników zapytania można również zmieniać przez interfejs użytkownika w Wikidata Query Service – wystarczy w polu wyników w lewej górnej części wybrać odpowiednią wizualizację. Na przykład na podstawie zapytania **4.9** można wygenerować diagram bąbelkowy pokazany na rys. **4.20**.

Istnieje również możliwość automatycznego generowania kodu do implementacji wyników zapytań w aplikacje napisane w różnych językach: PHP, Python, Perl, Java, JavaScript, HTML, Ruby, R, Matlab oraz innych.

Inne narzędzia do wizualizacji wyników zapytań w Wikidanych opisane na stronie https://www.wikidata.org/wiki/Wikidata:Tools/Visualize_data

RYSUNEK 4.19. **MAPA ZBUDOWANA NA PODSTAWIE ZAPYTANIA 4.10**

Źródło: https://query.wikidata.org

4.6.6. Łączenie danych w arkuszu kalkulacyjnym

RYSUNEK 4.20. **DIAGRAM BĄBELKOWY NA PODSTAWIE ZAPYTANIA 4.9 (Z LEWEJ STRONY POKAZANE RÓŻNE MOŻLIWOŚCI WIZUALIZACJI OTRZYMANYCH WYNIKÓW)**

Źródło: http://tinyurl.com/y4lowjwt

Arkusze kalkulacyjnych Google pozwalają korzystać z dodatkowych wtyczek, które mogą pomóc we wzbogaceniu istniejących danych w tabelach. W niniejszej sekcji rozpatrzymy dodatek „Wikipedia and Wikidata Tools", który umożliwia automatyczne pobieranie danych z Wikipedii oraz Wikidanych. W celu włączenia dodatkowych możliwości należy przejść w menu „Dodatki" > „Pobierz dodatki" > wpisać do wyszukiwarki nazwę wtyczki.

Lista wszystkich możliwych poleceń w ramach rozpatrywanego dodatku dostępna w menu „Dotatki" > „Wikipedia and Wikidata Tools" > „Show documentation". Rozpatrzymy niektóre z tych poleceń:

- WIKIGEOCOORDINATES – zwraca współrzędne geograficzne przypisane do artykułu z Wikipedii, tj. długość oraz szerokość geograficzną.
- WIKIPAGEVIEWS – zwraca statystyki dotyczące liczby odwiedzania stron Wikipedii.
- WIKIDATAQID - zwraca identyfikator dla odpowiadającego elementu Wikidanych.
- WIKITRANSLATE – zwraca tłumaczenia artykułu z Wikipedii.

Na przykład mamy tabelę wejściową z jedną kolumną o nazwie „Miasto", w której będzie wpisane 5 polskich miast (w każdej linijce jedno miasto): Warszawa, Kraków, Łódź, Wrocław, Poznań. Zatem mamy tylko 1 kolumnę oraz 6 wierszy (razem z nagłówkiem tabeli). Naszym celem jest automatyczne dodanie informacji o współrzędnych geograficznych (długość i szerokość), popularności (liczba odwiedzin za 2018 rok), odpowiednim identyfikatorze w Wikidanych oraz tłumaczenie nazwy w języku niemieckim dla każdego miasta. W każdym z rozpatrywanych poleceń dodatku „Wikipedia and Wikidata Tools", używany jest m.in. argument z nazwą artykułu, w którym musimy również wpisać wersję językową Wikipedii. Np. dla artykułu o Warszawie w polskojęzycznej Wikipedii nazwa będzie „pl:Warszawa". W celu otrzymania dodatkowych danych dla tej nazwy przy pomocy czterech rozpatrywanych poleceń z dodatku „Wikipedia and Wikidata Tools" należy w poszczególnych komórkach wpisać formuły umieszczone w listingu **4.11**

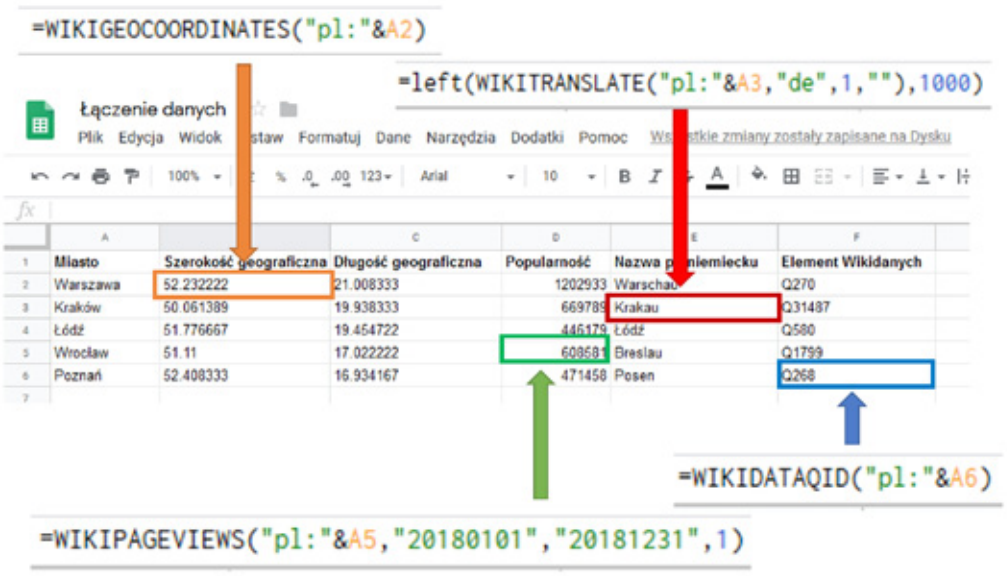
KOD ŹRÓDŁOWY 4.11. **POLECENIA ZWRACAJĄCE 5 DODATKOWYCH KOLUMN DLA ARTYKUŁU O WARSZAWIE W POLSKOJĘZYCZNEJ WIKIPEDII**

```
1    =WIKIGEOCOORDINATES("pl:Warszawa")
2    =WIKIPAGEVIEWS("pl:Warszawa","20180101","20181231",1)
3    =left(WIKITRANSLATE("pl:Warszawa","de",1,""),1000)
4    =WIKIDATAQID("pl:Warszawa")
```

W związku z tym, że nazwa artykułu zależy od wartości komórek z pierwszej kolumny w naszej pierwotnej tabeli, zamienimy nazwę do postaci „pl:&A2" dla komórki w drugim wierszu (w pierwszym wierszu podane są nagłówki tabeli). Te zmiany należy wykonać tylko dla jednego rekordu, po czym można skopiować wartości do innych wierszy (przy tym odwołania będą ustawione do odpowiedniej komórki z pierwszej kolumny z nazwami).

Wynikowa tabela wraz przykładami kodu dla różnych kolumn pokazana na rys. **4.21**.

RYSUNEK 4.21. **WYNIKOWA TABELA WRAZ PRZYKŁADAMI KODU ZDODATKU „WIKIPEDIA AND WIKIDATA TOOLS" DLA RÓŻNYCH KOLUMN W ARKUSZU KALKULACYJNYM GOOGLE**



Źródło: Opracowano na podstawie https://docs.google.com

5 Open Innovation – firmy dla społeczności

Niniejszy rozdział przedstawia działania w obszarze otwartych innowacji (ang. *open innovation*). Przedstawiono w nim koncepcję otwierania procesu innowacji oraz wskazano na różnice w stosunku od tradycyjnego podejścia do opracowywania innowacji.

Analiza źródeł wskazuje, że jednym z najbardziej rozpowszechnionych sposobów pozyskiwania otwartych innowacji są konkursy innowacyjności. Zagadnienie to, wraz z przykładami, zostało omówione w sekcji **5.2**. Sekcja **5.3** przedstawia rozwiązania skoncentrowane na wymianie gotowych algorytmów pomiędzy przedsiębiorstwami a społecznością. Ostatni obszar otwartych innowacji to wykorzystanie ich mechanizmów w analizie porównawczej – rozwiązania z tego zakresu omawia sekcja **5.4**. Rozdział kończy się analizą przypadków (sekcja **5.5**).

5.1. Wprowadzenie

Otwarte innowacje (ang. *open innovation*) to koncepcja mówiąca o tym, że w celu opracowywania innowacji przedsiębiorstwa nie muszą korzystać wyłącznie z własnych zasobów badawczo-rozwojowych, ale mogą uzupełniać je zasobami zewnętrznymi. Takimi zewnętrznymi zasobami mogą być inne podmioty biznesowe lub społeczność skupiona wokół określonego projektu lub portalu.

Open Innovation w założeniu opiera się na dwukierunkowej wymianie innowacji. Z jednej strony, przedsiębiorstwa korzystają z zasobów zewnętrznych (np. pozyskując wiedzę lub konkretne rozwiązania technologiczne), a z drugiej udostępniają na rynku te ze swoich rozwiązań, których nie wykorzystują. Zasady udostępnienia pozostają kwestią otwartą: może to być np. licencja (otwarta lub komercyjna), spin-off lub budowanie konsorcjum. Zasady pozyskiwania oraz udostępniania rozwiązań są jednym z elementów modelu biznesowego przedsiębiorstwa.

Termin „open innovation" został wprowadzony w 2003 roku przez Henry'ego Chesbrougha [**30**]. Wskazuje on na różnice pomiędzy tradycyjnym podejściem (nazywanym przez niego zamkniętymi innowacjami), a podejściem otwartym. Różnice te zebrane są w tabeli **5.1**. Otwarte innowacje pozwalają na czerpanie z wiedzy, doświadczenia i pomysłów ekspertów znajdujących się poza ekosystemem przedsiębiorstwa. Ponadto umożliwiają one wykorzystanie na wewnętrzne potrzeby wyników prac B+R, które zostały przeprowadzone przez zewnętrzne podmioty. Warto jednak podkreślić różnicę pomiędzy outsourcingiem usług B+R, w których całość prac jest zlecana innemu podmiotowi, a podejściem otwartych innowacji, gdzie wartość jest wypracowywana poprzez połączenie wyników prac B+R z zewnętrznych źródeł z wynikami wewnętrznymi przedsiębiorstwa.

TABELA 5.1. **PORÓWNANIE ZAMKNIĘTYCH I OTWARTYCH INNOWACJI**

ZASADY „ZAMKNIĘTYCH" INNOWACJI	ZASADY „OTWARTYCH" INNOWACJI
Eksperci pracują w danym przedsiębiorstwie.	Sięganie do wiedzy ekspertów spoza przedsiębiorstwa.
Żeby czerpać zyski z prac B+R, konieczne jest samodzielne odkrycie, opracowanie, wyprodukowanie i wprowadzenie na rynek innowacji.	Zewnętrzne B+R mogą dostarczyć wartość, choć konieczne są wewnętrzne działania B+R, żeby ją osiągnąć.
Kto pierwszy opracuje innowacje, pierwszy wprowadzi ją na rynek.	Nie trzeba być odkrywcą innowacji, żeby uzyskać z niej zyski.
Zwycięzcą jest ten, kto pierwszy skomercjalizuje innowację.	Lepiej jest opracować lepszy model biznesowy, niż być pierwszym na rynku.
Zwycięzcą jest ten, kto ma więcej lepszych pomysłów w branży.	Zwycięzcą jest ten, kto najlepiej wykorzysta wewnętrzne i zewnętrzne rozwiązania.
Należy kontrolować swoją własność intelektualną, aby inni nie mogli z niej czerpać.	Należy korzystać z własności intelektualnej innych, kupując te rozwiązania, które mogą rozwinąć model biznesowy.

Źródło: opracowanie na podstawie [30]

Istotnym elementem open innovation jest również związek pomiędzy innowacją a zbudowanym w oparciu o nią modelem biznesowym. Samo wprowadzenie innowacji na rynek niekoniecznie przyczynia się do uzyskania przewagi konkurencyjnej. Dopiero sprzężenie innowacji z właściwym modelem biznesowym pozwala na pełne czerpanie korzyści i budowanie przewagi konkurencyjnej.

Model biznesowy to narzędzie koncepcyjne, w którym elementy i powiązania między nimi pozwalają na przedstawienie logiki osiągania zysku przez przedsiębiorstwo. Jest to opis wartości, jaką przedsiębiorstwo dostarcza różnym grupom klientów, architektury przedsiębiorstwa oraz miejsca w sieci partnerów biznesowych, uwzględniając strumień kosztów i przychodów [**137**].

Hjalmarsson, Juell-Skielse i Johannesson [**75**] wskazują na cztery podstawowe modele otwartych innowacji:

- **Crowdsourcing** – outsourcing zadań do niezidentyfikowanej, zwykle szerokiej grupy ludzi, organizowany w formie otwartego ogłoszenia. W przypadku outsourcingu zadań związanych z innowacjami można ograniczyć grupę wykonawców, np. do klientów danej firmy, ale nie jest to wymagane. Crowdsourcing upowszechnił się wraz z pojawieniem się platform społecznościowych w Internecie.
- **Platformy produktowe** (ang. *product platform*) – zbiór narzędzi i zasobów, które są skonfigurowane i gotowe do wykorzystania, modyfikacji czy rozwijania przez użytkowników. Przykładem są platformy programistyczne (SDK – Software Development Kits). Platformy produktowe dostarczają wspólnej warstwy narzędziowo-pojęciowej dla organizacji, umożliwiając współpracę nie tylko wewnątrz organizacji, ale również włączając do procesów opracowywania innowacji zewnętrzne podmioty.
- **Innowacyjne sieci współpracy** (ang. *collaborative innovative network*) – zespół ambitnych osób, kierujących się wspólną wizją, współpracujących zdalnie poprzez Internet nad osiągnięciem wspólnego celu, wymieniających pomysły i informacje. Istotną charakterystyką takiej sieci jest brak hierarchii. Organizacje będące uczest-

nikami innowacyjnych sieci współpracy mogą dwojako z nich korzystać: czerpać inspiracje i rozwiązania od uczestników sieci, jak również udostępniać własne.
- **Konkursy innowacyjności** (ang. *innovative contest*) – konkursy, w których zadaniem uczestników jest opracowanie innowacji (idei, prototypu, produktu, usługi, metody, itp.). Zwykle oferowane są w nich stosunkowo wysokie nagrody pieniężne.

W praktyce najczęściej spotyka się modele mieszane, które łączą wszystkie (bądź większość) wyżej przedstawionych rozwiązań: zrzeszają społeczności innowacyjne, dostarczają zasobów danych i narzędzi na potrzeby użytkowników, jak również dostarczają rozwiązania techniczne umożliwiające przeprowadzanie konkursów innowacyjnych. Wiele organizacji zbudowało swój model biznesowy wokół organizacji konkursów innowacyjności – zagadnienie to zostało szerzej omówione w dalszej części rozdziału.

5.2. Konkursy innowacyjności i społeczności innowacyjne

Coraz popularniejszą metodą pozyskiwania innowacji stają się konkursy skierowane do społeczności innowacyjnych. Takie konkursy są prowadzone na portalach, które skupiają społeczności składające się zarówno z ekspertów, jak i osób dopiero zdobywających doświadczenie w określonej dziedzinie. Firma ogłaszająca konkurs definiuje problem badawczy, określa format rozwiązania (może to być model, algorytm, metoda, analiza czy gotowy prototyp), termin nadsyłania zgłoszeń oraz, opcjonalnie, nagrody finansowe. Najpopularniejsze platformy przyciągają nawet kilka tysięcy zespołów zgłaszających rozwiązania.

Podmioty powiązane z konkursami innowacyjności mogą występować w czterech rolach [**73**]:

- organizatorzy – obsługujący przebieg konkursu od strony technicznej i formalnej, zapewniający uczestnikom dostęp do zasobów, zarządzający infrastrukturą konkursową;
- uczestnicy – opracowujący rozwiązania, mogą występować jako uczestnicy indywidualni lub zespoły;
- dostawcyzasobów–dostawcydanych, narzędzi, zasobów sprzętowych, ale również środki finansowe lub know-how;
- beneficjenci – jednostki, które będą korzystać z wyników opracowanych przez uczestników.

Organizacja może występować w konkursie w kilku rolach jednocześnie, np. może być równocześnie organizatorem, dostawcą zasobów i beneficjentem. Na rynku dostępne są jednak dedykowane platformy do organizowania konkursów, które pozwalają wyraźnie rozdzielić te role. Idea otwartych innowacji zakłada, że każda z ról czerpie określone korzyści z konkursów. Opis korzyści osiąganych przez poszczególne role został przedstawiony w tabeli **5.2**.

TABELA 5.2. **ZADANIA I KORZYŚCI POSZCZEGÓLNYCH RÓL W KONKURSACH INNOWACYJNOŚCI**

ROLA	ZADANIA	KORZYŚCI
Organizator	Obsługa techniczna konkursu, zapewnienie platformy konkursowej, wsparcie merytoryczne dla beneficjenta, współpraca z beneficjentem przy definiowaniu zadań konkursowych.	Budowa społeczności, budowa wizerunku, w przypadku organizatorów działających komercyjnie: przychód z opłat za organizację.
Uczestnik	Opracowanie rozwiązania zadania konkursowego.	Budowawizerunku/reputacji, możliwość zdobycia nagród pieniężnych i rzeczowych, zdobywanie doświadczenie, nawiązywanie kontaktów (z innymi uczestnikami w ramach pracy zespołowej oraz z firmami-beneficjentami).
Dostawca zasobów	Dostarczenie narzędzi i danych na potrzeby realizacji zadań konkursowych, zapewnienie nagród, dostarczenie materiałów edukacyjnych.	Budowanie wizerunku, budowanie społeczności wokół własnych narzędzi, edukacja w zakresie wykorzystania własnych narzędzi.
Beneficjent	Definiowanie zadań konkursowych, ocena rozwiązań, wykorzystanie wyników konkursu.	Pozyskanie innowacyjnych rozwiązań, budowanie wizerunku, pozyskanie talentów spoza organizacji, nawiązanie kontaktów biznesowych.

Źródło: opracowanie własne.

Przedsiębiorstwa mogą czerpać korzyści z konkursów występując w każdej z ról. Z jednej strony, mogą występować jako beneficjent, pozyskując rozwiązania określonych problemów, z którymi się borykają. Takie podejście pozwala w efektywny sposób dotrzeć do dużej liczby potencjalnych ekspertów w dziedzinie, bez konieczności ograniczania się do swoich wewnętrznych zasobów. Warunki konkursów mogą być skonstruowane w taki sposób, aby prawo do komercyjnego wykorzystania rezultatów zwycięskich projektów przechodziło na organizatora. Co więcej, konkursy innowacyjności przyczyniają się do aktywizacji innowacyjnych sieci współpracy i społeczności skupionych wokół organizacji.

Z drugiej strony, przedsiębiorstwa, szczególnie start-upy, mogą wstępować jako uczestnik konkursu. Pozwala im to na pokazanie swojego potencjału, zmierzenie się z konkurencją oraz nawiązanie kontaktów biznesowych z większym partnerem, gdyż organizatorem konkursów są często największe firmy, światowi liderzy z różnych branż (np. Netflix, NASA, Microsoft, Alibaba Group).

Firmy wcielające się w rolę organizatorów, poza organizacją konkursów na własne potrzeby, mogą również świadczyć usługę w tym zakresie na rzecz innych podmiotów. Wokół tej roli swój model biznesowy zbudowały takie firmy jak Kaggle lub Tianchi.

Wreszcie, firmy mogą występować w roli dostawcy zasobów. Korzyści z tej roli to przede wszystkim upowszechnienie własnych technologii, budowanie wizerunku oraz budowanie społeczności.

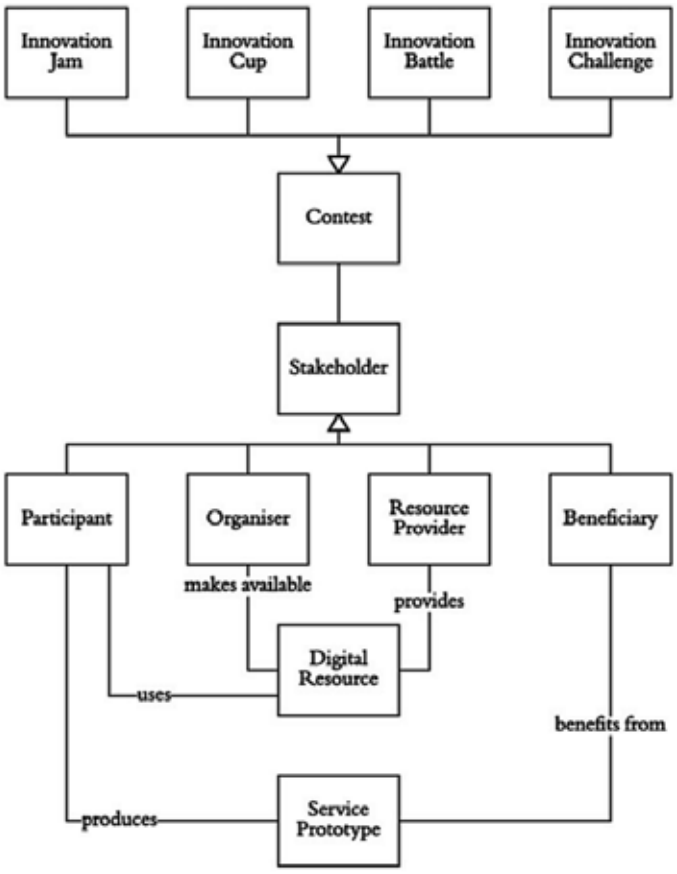
Hjalmarsson, Juell-Skielse i Johannesson [**73**] wyróżniają cztery typu konkursów innowacyjności:

- **Innovative Jam** – krótki czas trwania, skierowany do zamkniętej grupy uczestników, służący promocji określonego rozwiązania oraz stymulacji środowiska;

- **Innovative Battle** – krótki czas trwania, otwarty dla wszystkich chętnych, służący przede wszystkich sprawdzeniu możliwości realizacji jakiegoś pomysłu;
- **Innovation Cup** – długi czas trwania, skierowany do zamkniętej grupy uczestników, służący wzmacnianiu środowiska i rozwijaniu potencjału innowacyjnej sieci współpracy;
- **Innovation Challenge** – długi czas trwania, otwarty dla wszystkich chętnych, służący opracowaniu nowej idei, generowaniu innowacji.

Rysunek 5.1. przedstawia powiązania pomiędzy rolami oraz typami konkursów innowacyjności.

RYSUNEK 5.1. **PODSTAWOWE POJĘCIA ZWIĄZANE Z KONKURSAMI INNOWACYJNOŚCI**



Źródło: [73]

W dalszej części rozdziału przedstawiono przegląd platform organizujących konkursy innowacyjności. Część z nich jest pośrednikami (np. Kaggle), skupiającymi określoną społeczność i umożliwiającymi organizację konkursów skierowane do tej społeczności przez zewnętrzne podmioty. Inne (np. Yelp) łączą rolę organizatora i beneficjanta, organizując konkursy wyłącznie na własne potrzeby. Niektóre z platform służą również budowaniu kompetencji oferując zasoby e-learningowe. Porównanie platform konkursowych zawiera tabela **5.3**.

TABELA 5.3. **PORÓWNANIE KONKURSÓW INNOWACYJNOŚCI**

NAZWA	OBSZAR	CHARAKTER	LICZBA KONKURSÓW	NAGRODY KONKURSÓW[1] (W TYS. $)	LICZBA UCZESTNIKÓW[2]	INNE ELEMENTY
Kaggle	Machine learning, data science	Komercyjny, edukacyjny	19/295 (plus 4850 prywatnych)	0-1000	do 3000	Duży zbiór kursów z data science
DrivenData	Data science	Cele społeczne	6/15	0-50	Do 1300	
Crop Challenge		Komercyjny	1/3	5, 2,5, 1		
Tianchu	Big Data	Komercyjn	4/b.d.	0[3]	do 13000 zespołów	Tak, AI, po chińsku
CrowdAnalytix	AI	Komercyjny, edukacyjny	3/>120	0-10	Do 1000	forum
Innocentive	Bez ograniczeń	komercyjny	13/60	0-250	Do 500	Dodatkowe treści: webinary, white papers
Yels	NLP, przetwrzanie obrazu, graph ining	Edukacyjny	1/11[4]	10x 5	b.d.	
TunelT	Data mining	Komercyjny, edukacyjny	komercyjne 2/10, studenckie 3/22	0-45	Do 820	Repozytorium algorytmów
The Best Student VW	różny	Komercyjny	1/7[5]	2,6/0,8/ 0,25[6]	b.d.	

Źródło: opracowanie własne.

5.2.1. Kaggle

Kaggle[7] to najpopularniejsza i największa platforma konkursów z dziedziny data science. Celem konkursów jest najczęściej opracowanie modelu predykcyjnego na podstawie zbioru danych udostępnionych przez organizatora. Konkursy organizowane są przez firmy lub organizacje pozarządowe. Na platformie dzielą się na komercyjne (z prze-

1 Aktywne/zamknięte, stan na dzień 30.01.2019 r.
2 Liczba uczestników w pojedynczym konkursie
3 Jest przewidziana możliwość nagród, ale konkursy aktywne na dzień 30.01.2019 nie mają określonych.
4 Kolejne rundy tego samego konkursu
5 W każdej edycji 3-6 tematów konkursowych.
6 Przyznawane dla każdego tematu, zwycięzcy dodatkowo otrzymują praktykę w firmie.
7 https://kaggle.com/competitions

widzianą nagrodą pieniężną), edukacyjne (bez nagród, ale często są częścią zaliczenia kursów na portalach e-learningowych, np. Coursera[8]) lub zamknięte (tzw. „in class", organizowane dla zamkniętej grupy studentów w ramach zajęć edukacyjnych). Ostatni typ konkursów jest najbardziej popularny, ale nastawiony na zastosowania niekomercyjne i ze względu na cel (niekomercyjny, czysto edukacyjny, nie związany z realnym problemem biznesowym) nie wpisuje się w otwarte innowacje. Większość konkursów skierowana jest do zespołów, ale są również takie, gdzie wymagane są zgłoszenia indywidualne. Ponadto strona oferuje bogate zasoby zbiorów danych z różnych dziedzin.

Typy konkursów, które mogą być organizowane na Kaggle, to:

- *In class* – przeznaczone dla nauczycieli akademickich, darmowe. Polegają na zdefiniowaniu problemu badawczego i zbioru danych. Możliwe jest określenie kryteriów oceny, automatyczne ocenianie prac, szeregowanie rozwiązań oraz obsługa forum dyskusyjnego dla uczestników.
- *Featured* – konkursy komercyjne, polegające na rozwiązaniu konkretnego problemu predykcyjnego, zdefiniowanego przez przedsiębiorstwo. Udział w tych konkursach biorą głównie eksperci z danej dziedziny, ponieważ są w nich zwykle oferowane wysokie pule nagród (nawet 1.000.000 USD). Dostęp jest jednak otwarty dla każdego. Konkursy z tej grupy są dobrym przykładem otwartych innowacji.
- *Research* – konkursy dotyczące problemów badawczych, eksperymentalnych. Zwykle nie mają nagród, ale pozwalają zdobyć doświadczenie uczestnikom. Co więcej, zdarza się, że postawione problemy nie mają rozwiązania.
- *Getting Started* – proste zadania, przeznaczone dla początkujących w dziedzinie uczenia maszynowego i analizy danych. Nie mają nagród. Ich czas trwania jest bardzo długi, więc ranking uczestników jest utrzymywany w systemie 2-miesięcznym (rozwiązanie jest brane pod uwagę w rankingu tylko przez 2 miesiące od jego zgłoszenia).
- *Playground* – zadania o stopień trudniejsze, niż *getting started*, jednak wciąż przeznaczone dla początkujących. Większość dotyczy prostych problemów uczenia maszynowego.
- *Recruitment* – konkursy komercyjne, w których uczestnicy startują wyłącznie samodzielnie (zespoły jednoosobowe). Po zamknięciu konkursu uczestnicy mogą przesłać organizatorowi swoje CV, a nagrodą jest rozmowa rekrutacyjna.
- *Annual* – doroczne konkursy organizowane przez Kaggle: w marcu i grudniu.
- *Limited participation* – konkursy organizowane sporadycznie, przeznaczone dla wybranej, zamkniętej grupy uczestników.

8 https://www.coursera.org

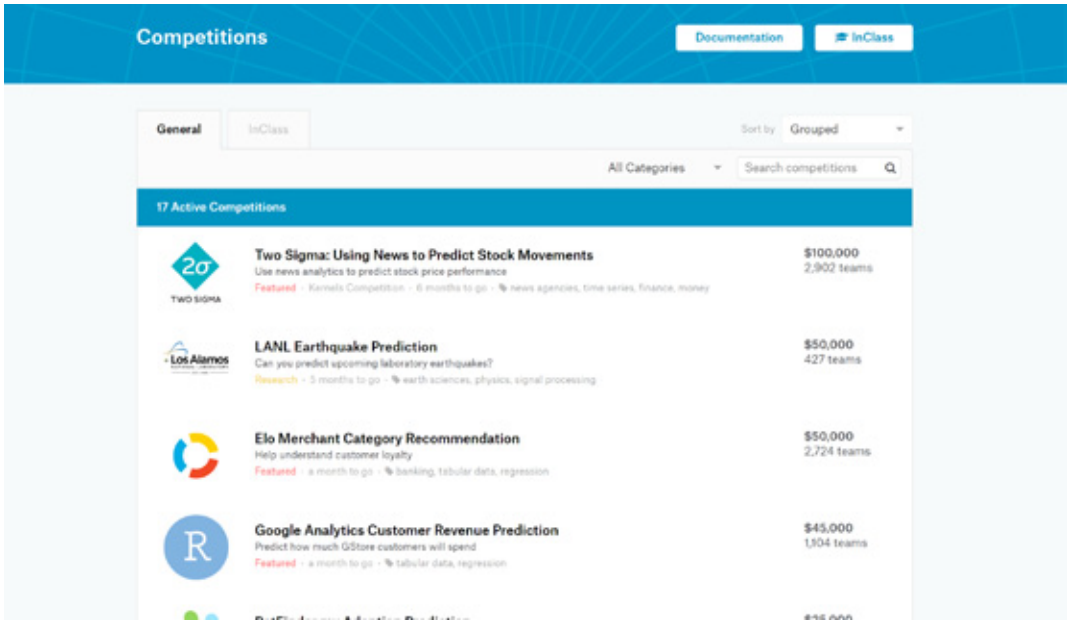
Kaggle oferuje organizację konkursów w kilku formatach:

- *Simple (classic)* – po zaakceptowaniu zasad konkursu uczestnicy mają pełny dostęp do danych konkursowych. Uczestnik pobiera dane, opracowuje model lokalnie lub bezpośredniona platformie (tzw.*kernel*) i jako rozwiązanie przesyła plik z modelem prognostycznym.
- *Two-stage* – w drugim etapie uczestnicy otrzymują nowy zbiór danych i wykorzystują rozwiązanie stworzone w etapie 1. Uczestnictwo w 2. etapie jest możliwe tylko pod warunkiem ukończenia etapu 1. Wcześniej dane etapu 2. są niedostępne.
- *Kernels-only* – tzw. konkursy kodu, wszystkie zgłoszenia tworzone są bezpośrednio na platformie Kaggle Kernels. Wszyscy uczestnicy działają na tym samym sprzęcie (Kaggle). Zwykle zwycięskie modele są mniej rozbudowane niż w innych formatach, gdyż muszą spełniać ograniczenia sprzętowe narzucane przez Kaggle.

Przykładowa lista konkursów na platformie Kaggle jest przedstawiona na rys. **5.2**. Z punktu widzenia pozyskiwania innowacji dla przedsiębiorstwa największe znaczenia mają konkursu typu *featured*, pozwalające zdobyć wiele alternatywnych rozwiązań postawionego problemu, oraz *recruitment*, służące do pozyskiwania kapitału ludzkiego.

Wśród firm korzystających z konkursów innowacyjności na Kaggle, są m.in.AllState, BNP Paribas, CAT, Facebook, Liberty Mutual, Merck, Santander oraz Microsoft.

RYSUNEK 5.2. **KAGGLE – LISTA KONKURSÓW**



Źródło: https://www.kaggle.com/competitions

5.2.2. DrivenData

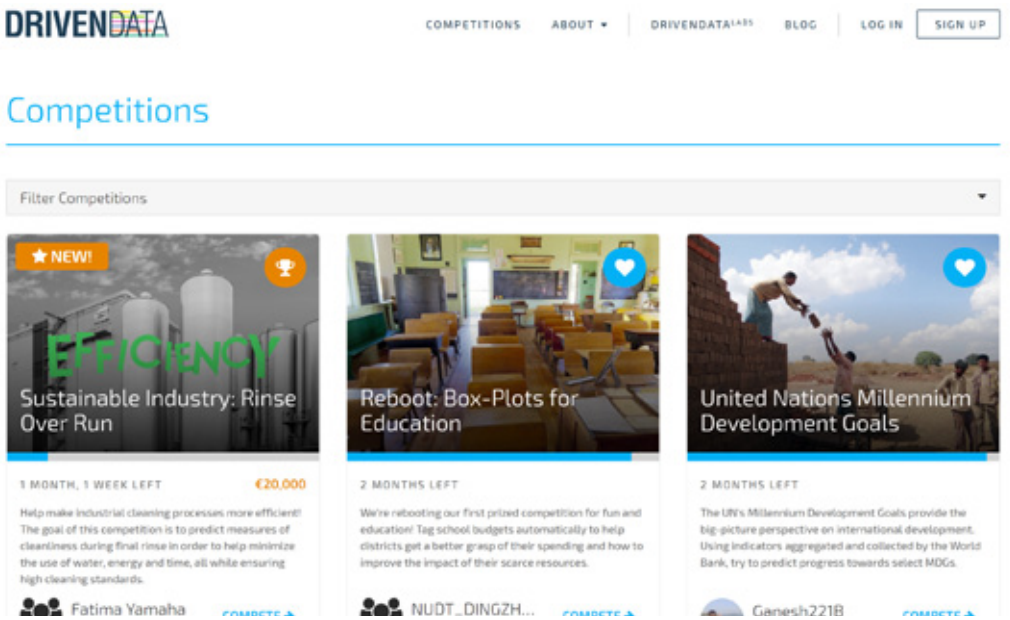
Platforma Driven Data[9] działa na podobnych zasadach jak Kaggle, ale problemy konkursowe dotyczą zagadnień zgłaszanych przez organizacje non-profit.

Celem konkursów jest znalezienie rozwiązanie określonych problemów społecznych, poprzez wykorzystanie idei crowdsourcingu. Konkursy trwają ok. 2-3 miesięcy, a zadaniem uczestników jest zbudowanie określonego modelu predykcyjnego. Zadania konkursowe są formułowane wspólnie z organizacjami non-profit. Zwycięskie rozwiązania są implementowane w tych organizacjach i testowane na zbiorach rzeczywistych danych.

Platforma oferuje konkursy kilku typów (rys. **5.3**):

- o określonym poziomie trudności: dla początkujących, średnio-zaawansowanych, zaawansowanych;
- z nagrodą pieniężną;
- edukacyjne (nazywane na platformie „for fun").

RYSUNEK 5.3. **DRIVENDATA – LISTA KONKURSÓW**



Źródło: https://www.drivendata.org/, dostęp 30.01.2019 r.

9 https://www.drivendata.org/

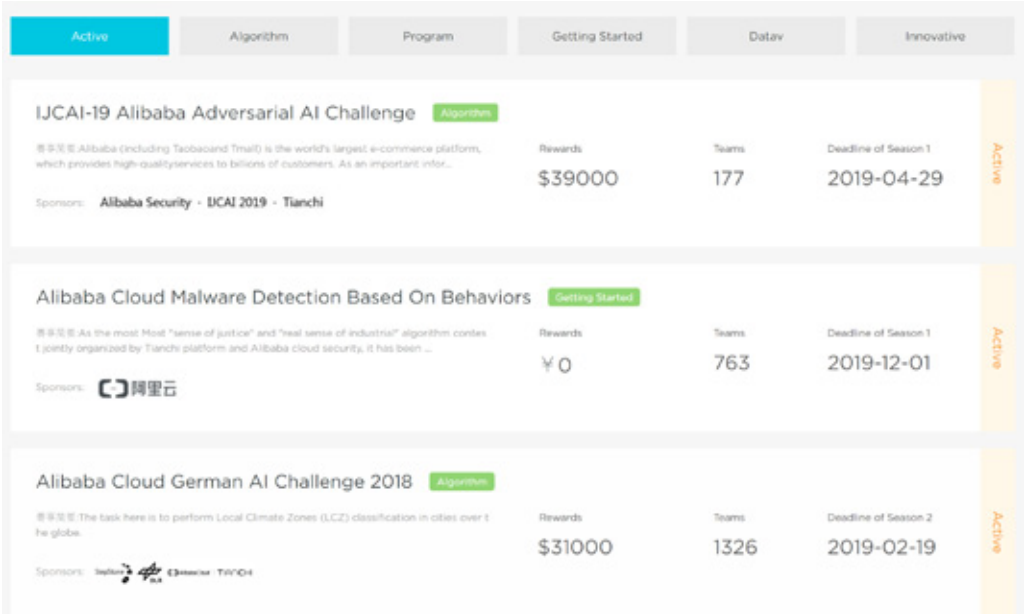
5.2.3. Tianchi

Tianchi[10] to platforma firmy będącej dostawcą technologii big data oraz zasobów obliczeniowych w formie usługi. Jest częścią Alibaba Cloud. Zasady działania Tianchi są bardzo zbliżone do Kaggle.

Konkursy mogą mieć jeden z kilku dostępnych typów (Algorithm, Program, Data, Innovative), jednak strona nie dostarcza otwartej dokumentacji określającej, jakie są charakterystyki poszczególnych konkursów. Wadą serwisu jest fakt, że duża część treści jest dostępna wyłącznie w języku chińskim, również część dokumentacji (np. pomoc). Ponadto nie ma wglądu w zamknięte konkursy.

Platforma konkursów innowacyjności jest tylko uzupełnieniem oferty usług e-learningowych z obszaru AI (kursy są dostępne w języku chińskim). Jak widać, strona kierowana jest głównie do społeczności chińskiej, ale gromadzi ponad 150 tys. użytkowników z ponad 80 krajów. Zawiera również zbiory danych. Ze względu na fakt, że Tianchi należy do dużej grupy finansowej, możliwy jest rozwój platformy w najbliższych latach.

RYSUNEK 5.4. **TIANCHI – LISTA KONKURSÓW**



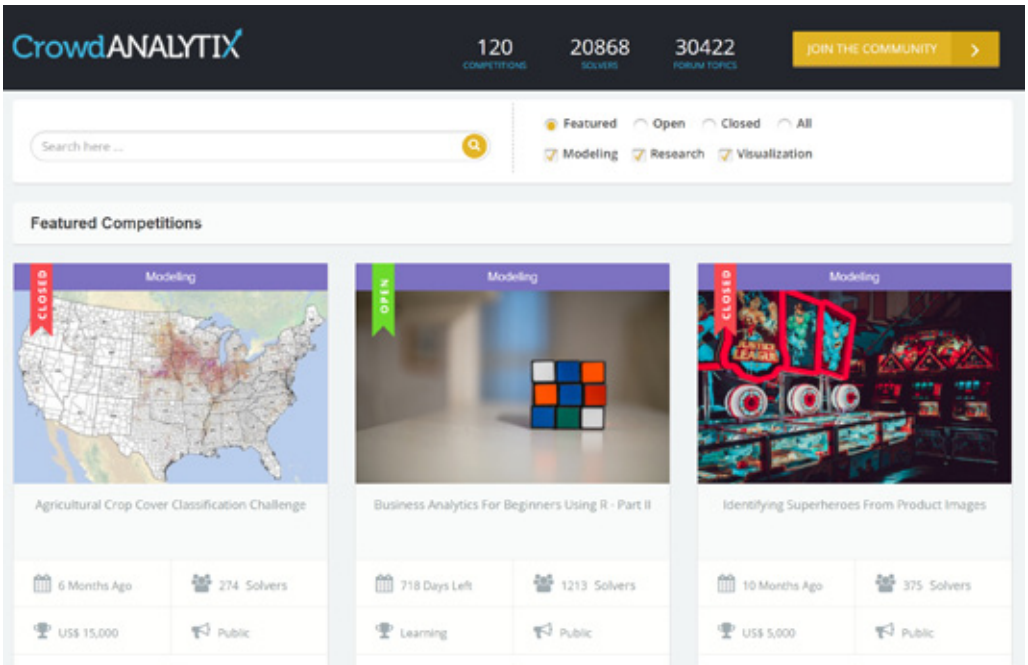
Źródło: https://tianchi.aliyun.com/competition/gameList/activeList, dostęp30.01.2019 r.

10 https://tianchi.aliyun.com/competition/gameList/activeList

5.2.4. CrowdAnalytix

Platforma CrowdAnalytix[11] wykorzystuje model crowdsourcingu – społeczność opracowuje rozwiązania postawionych problemów, dotyczących głównie zagadnień z dziedziny sztucznej inteligencji (AI). Społeczność liczy ponad 20 tys. użytkowników. Platforma zawiera bibliotekę botów, które potrafią ekstrahować informację z tekstu, obrazu, wideo. Konkursy innowacyjności na platformie podzielone są na trzy typy (rys. **5.5**): modelowanie (ang. *modeling*), badania (ang. *research*), wizualizacja (ang. *visualisation*).

RYSUNEK 5.5. **CROWDANALYTIX – LISTA KONKURSÓW**



Źródło: https://www.crowdanalytix.com/en/, dostęp 30.01.2019 r.

11 https://www.crowdanalytix.com/en/

5.2.5. Innocentive

Innocentive[12] zrzesza ponad 400 tys. użytkowników. Platforma nie stawia ograniczeń dziedzinowych dla organizowanych konkursów – mogą one dotyczyć dowolnego zagadnienia.

Innocentive pozwala na organizację następujących rodzajów wyzwań (konkursów):

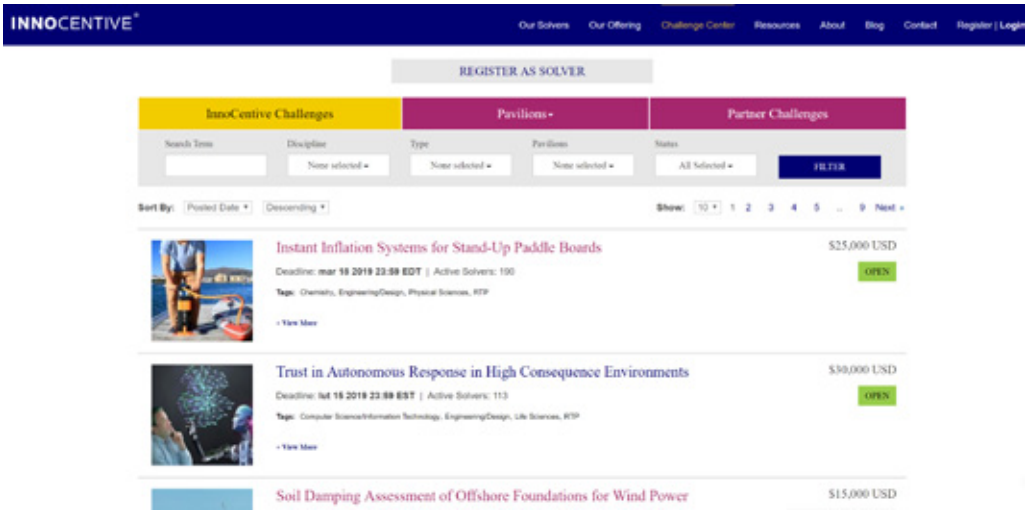
- *Ideation Challenges* – „burze mózgów", ich celem nie jest znalezienie konkretnego rozwiązania, a raczej wygenerowanie przestrzeni możliwych rozwiązań.
- *Theoretical Challenges* – celem jest opracowanie konkretnego rozwiązania teoretycznego, dostarczonego w formie raportu.
- *Reduction to Practice (RTP) Challenges* – ocenie podlegają rozwiązania, które zostały przetestowane i sprawdzone eksperymentalnie.
- *Electronic Request for Partner (eRFP) Challenges* – konkurs służy poszukiwaniu partnerów biznesowych.
- *Showcase Challenges* – przeznaczony wyłacznie dla firm, start-upów z określonego obszaru biznesowego. Pozwala znaleźć partnera lub ciekawą aktywność w określonej branży
- *Prodigy „Big Data"* – rozszerzenie konkursów typu Theoretical i RTP, umożliwiające przekazanie uczestnikom informacji zwrotnej (porównania wyniku w stosunku do innych rozwiązujących).
- *Novel Molecule Challenge (NMC)* – zgłoszenie zapotrzebowania na niekomercyjne związki chemiczne, polimery, sekwencje DNA itp. Zwykle nagroda zależy od przekazanych praw: przy braku wyłączności jest mniejsza, większa jeżeli zostaną przekazane wyłączne prawa do wykorzystania.

Konkursy (wyzwania) dostępne na stronie dzielą się ponadto na „zwykłe" (otwarte, na stronie widać ok. 60) oraz „premium" (ponad 2000) (rys. **5.6**)

Platforma oferuje wsparcie dla twórców wyzwań w postaci organizacji warsztatu, na którym eksperci pomagają sformułować problem konkursowy.

12 https://www.innocentive.com/

RYSUNEK 5.6. **INNOCENTIVE – LISTA KONKURSÓW**



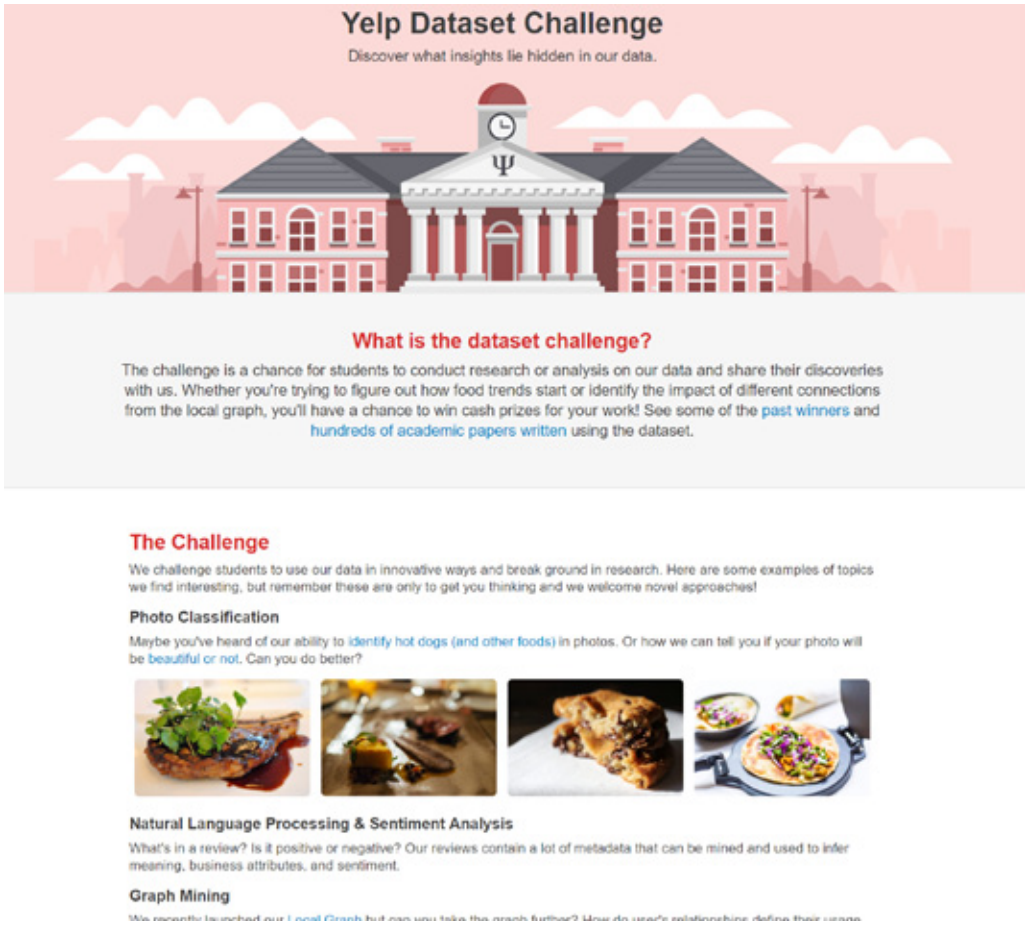
Źródło: https://www.innocentive.com/, dostęp 30.01.2019 r.

5.2.6. Yelp

Yelp[13] to cykliczny konkurs analizy danych, organizowany przez portal na własnych danych. Zbiór danych zawiera recenzje, obiekty biznesowe, zdjęcia i dane geograficzne. Kolejne rundy konkursowe dotyczą problematyki analizy dostarczanego zbioru, np. klasyfikacji zdjęć, przetwarzania języka naturalnego i analizy wydźwięku, eksploracji grafów. Konkursy skupiają się na eksploracji możliwych sposobów wykorzystania określonego, ograniczonego tematycznie zbioru danych.

13 https://www.yelp.com/dataset/challenge

RYSUNEK 5.7. **YELP – STRONA KONKURSU**



Źródło: https://www.yelp.com/dataset/challenge, dostęp 30.01.2019 r.

5.2.7. Idea Connection

Idea Connection[14] nie jest platformą konkursową, a pojedynczym konkursem organizowanym od 2016 roku. Zgłoszenia są otwarte dla wszystkich (pełnoletnich) chętnych osób. Konkurs organizowany jest przez The Analytics Society of INFORMS. Dane konkursowe nie są otwarte – dostęp do nich otrzymuje się po zaakceptowaniu warunków konkursu i nie można ich wykorzystywać ani upubliczniać w żadnym innym celu, niż udział w konkursie.

RYSUNEK 5.8. **IDEA CONNECTION – KONKURS OGŁOSZONY NA ROK 2019**

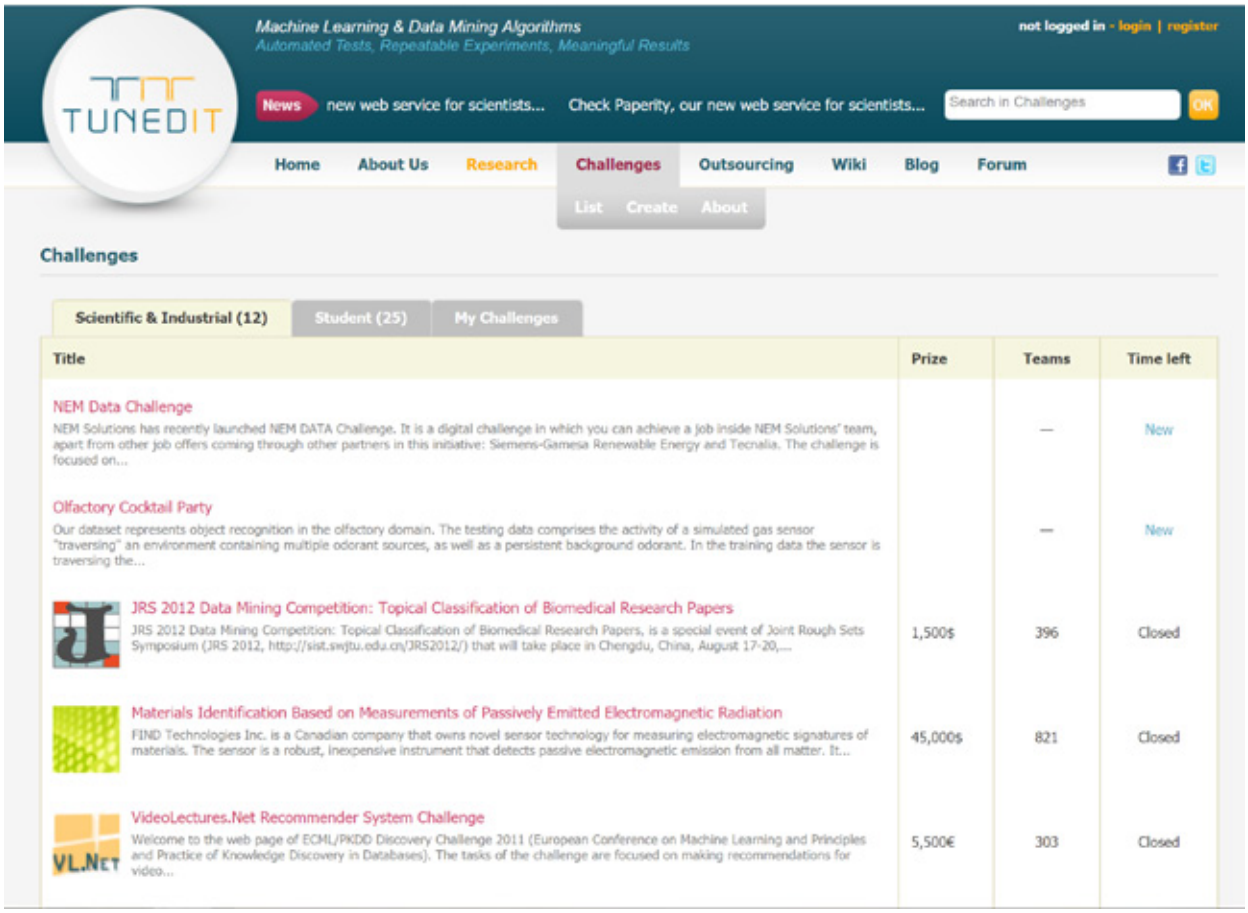
Źródło: https://www.ideaconnection.com/syngenta-crop-challenge/, dostęp 30.01.2019 r.

14 https://www.ideaconnection.com/syngenta-crop-challenge/

5.2.8. TuneIT

TuneIT[15] to polska platforma, założona jako projekt naukowy, a później przekształcona w spin--off. Oferuje darmowe narzędzia do analizy danych dla naukowców, a dla firm możliwość organizowana konkursów na opracowanie innowacyjnych algorytmów analizy danych (rys. **5.9**). Pozwala organizować konkursy studenckie, naukowe i przemysłowe (dwa ostatnie za opłatą). Zawiera również repozytorium zbiorów danych.

RYSUNEK 5.9. **TUNEIT – LISTA KONKURSÓW**



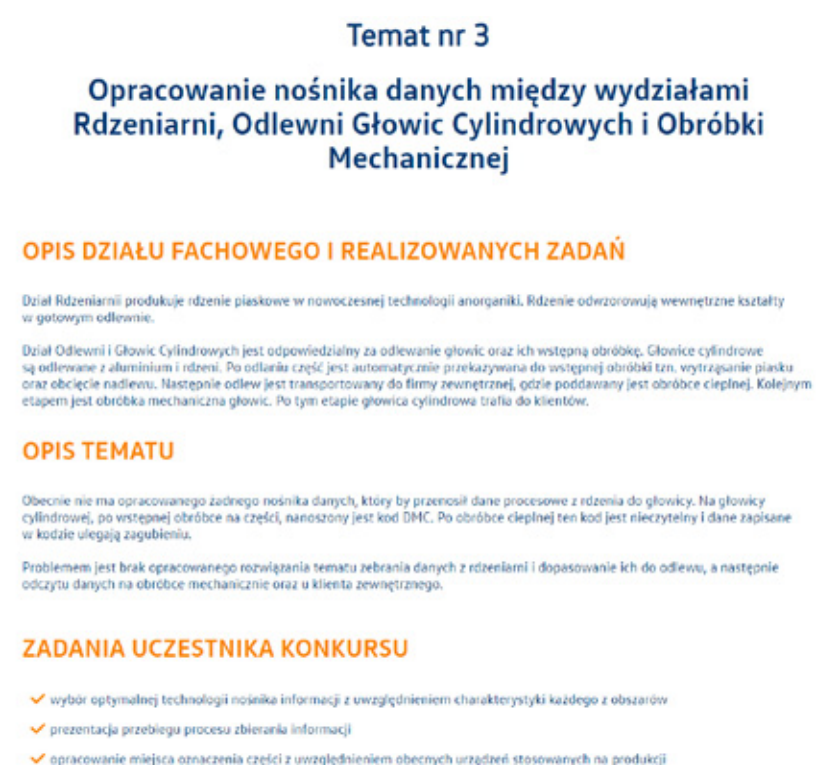
Źródło: http://tunedit.org/, dostęp 30.01.2019 r.

15 http://tunedit.org/

5.2.9. The Best Student VW

Cykl konkursów Volkswagen Poznań Sp. z o.o.[16] jest skierowany do studentów. Uczestnicy mogą startować indywidualnie lub w zespołach dwuosobowych, a ich zadaniem jest opracowanie projektu na realnym przypadku użycia, dostarczonym przez organizatorów konkursu. W każdej edycji publikowanych jest kilka różnych tematów projektowych (w roku 2019 jest ich 6). Tematy zawierają opis problemu oraz wymagania w stosunku do formatu rozwiązania (np. aplikacja mobilna na dany system operacyjny, koncept projektu rozwiązującego zadany problem, model 3D). Do każdego tematu konkursowego przypisani są też opiekunowie – przedstawiciele beneficjenta – którzy mają pomóc uczestnikom w opracowaniu finalnego rozwiązania. Dla każdego tematu wybieranych jest maksymalnie 15 zespołów, które przechodzą do etapu 2., gdzie pomysły są doprecyzowywane.

RYSUNEK 5.10. **PRZYKŁADOWY TEMAT W RAMACH KUNKURSU VW THE BEST STUDENT**



Temat nr 3

Opracowanie nośnika danych między wydziałami Rdzeniarni, Odlewni Głowic Cylindrowych i Obróbki Mechanicznej

OPIS DZIAŁU FACHOWEGO I REALIZOWANYCH ZADAŃ

Dział Rdzeniarnii produkuje rdzenie piaskowe w nowoczesnej technologii anorganiki. Rdzenie odwzorowują wewnętrzne kształty w gotowym odlewnie.

Dział Odlewni i Głowic Cylindrowych jest odpowiedzialny za odlewanie głowic oraz ich wstępną obróbkę. Głowice cylindrowe są odlewane z aluminium i rdzeni. Po odlaniu część jest automatycznie przekazywana do wstępnej obróbki tzn. wytrząsanie piasku oraz obcięcie nadlewu. Następnie odlew jest transportowany do firmy zewnętrznej, gdzie poddawany jest obróbce cieplnej. Kolejnym etapem jest obróbka mechaniczna głowic. Po tym etapie głowica cylindrowa trafia do klientów.

OPIS TEMATU

Obecnie nie ma opracowanego żadnego nośnika danych, który by przenosił dane procesowe z rdzenia do głowicy. Na głowicy cylindrowej, po wstępnej obróbce na części, nanoszony jest kod DMC. Po obróbce cieplnej ten kod jest nieczytelny i dane zapisane w kodzie ulegają zagubieniu.

Problemem jest brak opracowanego rozwiązania tematu zebrania danych z rdzeniarni i dopasowanie ich do odlewu, a następnie odczytu danych na obróbce mechanicznie oraz u klienta zewnętrznego.

ZADANIA UCZESTNIKA KONKURSU

- wybór optymalnej technologii nośnika informacji z uwzględnieniem charakterystyki każdego z obszarów
- prezentacja przebiegu procesu zbierania informacji
- opracowanie miejsca oznaczenia części z uwzględnieniem obecnych urządzeń stosowanych na produkcji

Źródło: http://www.bethebeststudent.pl/, dostęp 30.01.2019 r.

16 http://www.bethebeststudent.pl/

5.2.10. NumerAI

NumerAI[17] jest platformą konkursów innowacyjności, stosującą zupełnie inny model biznesowy niż pozostałe tego typu platformy. NumerAI wykorzystuje zanonimizowane dane z rynków finansowych, na których można przeprowadzać analizy, które nie będą obciążone przekonaniami człowieka lub informacjami z innych źródeł. Modele opracowane przez uczestników konkursów są wykorzystywane przez powiązany z platformą fundusz hedgingowy, a nagrody wypłacane są w dedykowanej kryptowalucie. Uczestnicy pracują na danych, które stanowią abstrakcyjną reprezentacją pewnego rynku finansowego (nie da się na ich podstawie określić, jakiego rynku dotyczą). Uczestnicy opracowują modele, które mają wykrywać wzorce w danych. NumerAI następnie na podstawie modeli użytkowników opracowuje własny meta-model, na podstawie którego podejmuje decyzje o swoich inwestycjach. 100 najlepszych modeli w każdym konkursie jest nagradzanych.

Konkursy mają formę cotygodniowych turniejów. Zgłoszone w każdym konkursie modele są porównywane między sobą i oceniane funkcją *Log Loss*. Uczestnik może „założyć się", że jego model będzie lepszy niż inne – wtedy ma szansę na wygranie nagród pieniężnych. „Zakład" musi zostać zgłoszony na określony czas przed zamknięciem konkursu – lista rankingowa rozwiązań jest aktualizowana na bieżąco, więc w ostatnich godzinach konkursu nie można już dołączyć do zakładów.

5.3. Otwarte algorytmy

Istotną barierą w rozwoju otwartych zbiorów danych jest konieczność zapewnienia prywatności i anonimizacji danych wrażliwych. Obowiązujące przepisy prawne wymagają, aby dane przed udostępnieniem były przekształcone w sposób, który uniemożliwi identyfikację pojedynczych obserwowanych obiektów. Osiąga się to poprzez usunięcie charakterystyk mogących zidentyfikować osoby lub poprzez opracowanie wartości zagregowanych. Przygotowanie danych w ten sposób jest pracochłonne i kosztowne, co ogranicza szanse na ich upublicznienie.

Projekt otwartych algorytmów OPAL podchodzi do kwestii prywatności w inny sposób: zamiast anonimizacji danych zakłada przetworzenie ich zadanym algorytmem na platformie właściciela danych i zwrócenie twórcy algorytmu jedynie gotowych wyników. Takie podejście umożliwia udostępnianie danych do przeprowadzenia określonych analiz, ale bez faktycznego ujawniania konkretnych rekordów czy przekazywania kopii całego zbioru.

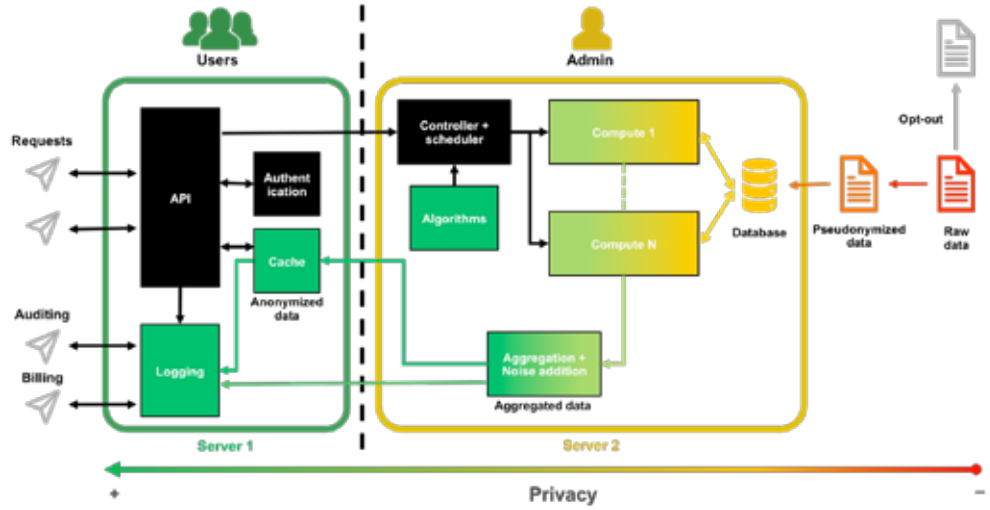
17 https://numer.ai/homepage

5.3.1. OPAL – Open Algorithms Project

Projekt OPAL[18], działający od 2017 roku, jest obecnie najpopularniejszą i najbardziej rozwiniętą inicjatywą związaną z otwartymi algorytmami. Projekt jest prowadzony przez konsorcjum złożone z MIT, Orange Group, Imperial Collage London, World Economic Forum, Data-Pop Alliance and MIT Media Lab oraz Overseas Development Institute. Do tej pory zostały zrealizowane dwa projekty pilotażowe o wartości 1,5 mln USD, przeprowadzone w Senegalu i Kolumbii. Wdrożenia w 10 kolejnych krajach planowane są do roku 2025.

OPAL to platforma typu non-profit. Jej celem jest umożliwienie aktywnego i efektywnego wykorzystania danych prywatnych do celów społecznych, z poszanowaniem prywatności. Ma to zapewniać odpowiednio zaprojektowana architektura, przedstawiona na rys. **5.11**. Jako jedno z zastosowań twórcy platformy wskazują możliwość wygenerowania o wiele dokładniejszych agregatów opisujących społeczności niż to ma miejsce w oficjalnych statystykach – takie agregaty byłyby tworzone w oparciu o rzeczywiste dane, zamiast szacunków instytucji statystycznych. Przykładem rzeczywistych danych mogą być dane firm telekomunikacyjnych, które przy wykorzystaniu technologii big data umożliwiają wnioskowanie o zachowaniach, preferencjach oraz wzajemnych relacjach ludzi, ale ze względów prawnych i etycznych ich wykorzystanie jest znacznie ograniczone. Tego typu dane nie są też udostępniane jako otwarte zbiory.

RYSUNEK 5.11. **ARCHITEKTURA PLATFORMY OPAL**



Źródło: https://www.opalproject.org/technical-component, dostęp 30.01.2019 r.

18 https://www.opalproject.org/

Problemy, które OPAL stara się rozwiązać, to:

- brak aktualnych, szczegółowych i rzetelnych statystyk i wskaźników, które mogłyby wspierać podejmowanie decyzji przez przedsiębiorstwa;
- uregulowania prawne, które uniemożliwiają wykorzystanie wielu wartościowych, szczegółowych zbiorów danych o wysokiej jakości, znajdujących się w posiadaniu firm prywatnych;
- brak systemowych rozwiązań umożliwiających wykorzystanie nowych źródeł danych i technik ich przetwarzania, szczególnie w krajach rozwijających się.

Cele projektu OPAL to:

- Opracowanie platformy łączącej sektor prywatny, publiczny oraz zwykłych obywateli (Public-Private-People), która przyczyni się do rozwoju społecznego, zgodnie z zasadami zrównoważonego rozwoju i demokracji.
- Dostarczenie lepszego opisu rzeczywistości niż oferowany przez urzędy i organizacje statystyczne, a tym samym wspieranie podejmowania decyzji we wszystkich sektorach.
- Umożliwienie zaangażowania członków społeczności w tworzenie i analizowanie danych o nich samych.

5.3.2. Algorithms Open Marketplace

Algorithmia jest platformą, której główna usługa polega na umożliwieniu „handlu" algorytmami. Model biznesowy Algorithmia[19] składa się z kilku elementów. Po pierwsze, platforma udostępnia algorytmy w formie usługi – możliwe jest wywołanie wybranego algorytmu we własnej aplikacji. Pozwala to na przeniesienie wykonania określonych algorytmów z aplikacji na zasoby Algorithmia, a tym samym poprawę wydajności, skalowalności i efektywności aplikacji. Z drugiej strony, integralnym elementem portalu jest społeczność zajmująca się uczeniem maszynowym – tworzone przez nią modele są dostępne w repozytorium i stanowią „ofertę" usług, które można wykorzystywać we własnych aplikacjach. Społeczność skupia ponad 60 tys. deweloperów, którzy opracowali i udostępnili już ponad 4500 algorytmów[20].

Portal kieruje swoją ofertę do trzech grup użytkowników:

- inżynierów tworzących algorytmy i oferujących je poprzez platformę,
- analityków danych, którzy mogę korzystać z algorytmów udostępnionych w formie usług w celu analizy własnych zbiorów danych;

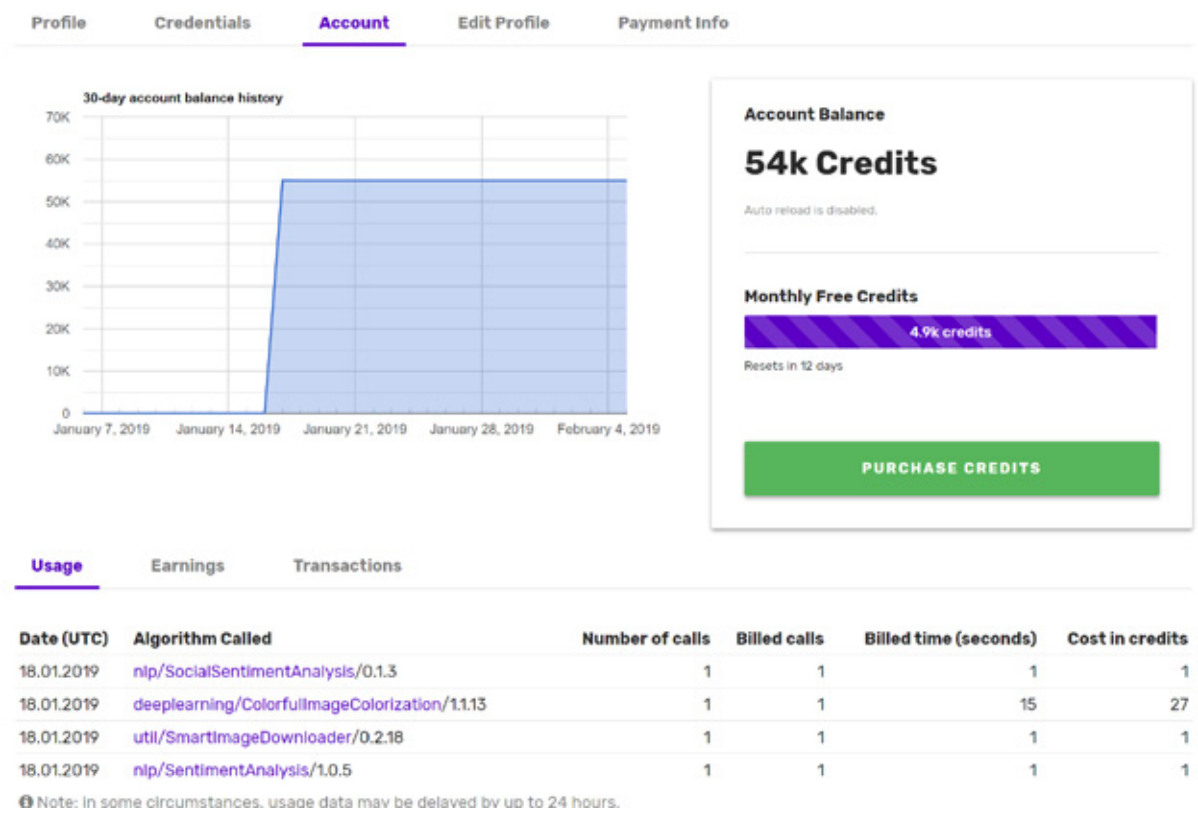
19 https://algorithmia.com/

20 https://www.forbes.com/sites/amitchowdhry/2018/01/22/how-algorithmia-built-the-largest-marketplacefor-algorithms-in-the-world/

- przedsiębiorstw, do których skierowana jest kompleksowa oferta zarządzania portfolio modeli uczenia maszynowego, utrzymywanego w prywatnej chmurze lub jako rozwiązanie on-premise.

Platforma wprowadziła też wewnętrzny sposób rozliczania usług, tzw. *kredyty* (ang. *credits*). Każdy użytkownik otrzymuje co miesiąc 5000 darmowych kredytów. Może je wykorzystać na wywoływanie usług („płacąc" 1 kredyt za 1 sekundę czasu obliczeniowego oraz opłaty licencyjne, jeżeli były określone przez twórcę algorytmu). Dodatkowe kredyty można dokupić (10.000 kredytów to 1 USD) lub zarobić wystawiając własne algorytmy na portalu. Kredyty zgromadzone dzięki wywołaniom usługi użytkownika przez innych można zamienić na gotówkę (minimalna wypłata to 100 USD). Panel użytkownika przedstawiony jest na rys. **5.12**.

RYSUNEK 5.12. **PORTAL ALGORITHMIA – PODSUMOWANIE KREDYTÓW UŻYTKOWNIKA**



Źródło: https://algorithmia.com/, dostęp 30.01.2019 r.

5.4. Benchmarking

Termin benchmarking odnosi się do zbioru metod analitycznych służących porównaniu określonego zjawiska lub charakterystyk pomiędzy co najmniej dwoma obiektami. Benchmarking jest wykorzystywany do określenia pozycji rynkowej przedsiębiorstwa względem konkurencji. Może być również metodą pozyskania wiedzy, np. poprzez porównanie własnych rozwiązań w zakresie sposobu przeprowadzenia określonych procesów biznesowych z rozwiązaniami konkurencji.

5.4.1. APQC – benchmarkig portal

Portal APQC[21] dostarcza danych o wydajności organizacji z określonych branż. Użytkownicy portalu mogą przeglądać dane benchmarkowe lub wypełnić jeden z oferowanych formularzy oceny efektywności, żeby otrzymać raport pozycjonujący przedsiębiorstwo na tle konkurencji.

Część zasobów portalu jest dostępna tylko dla organizacji, które uiściły opłatę członkowską. Płatne zasoby to przede wszystkim narzędzia benchmarkowe. Na portalu jest również dostępna baza wiedzy zawierająca opisy standardów i metod oraz inne materiały związane z analizą organizacji. Jest ona w większości otwarta. Co ważne, opłata członkowska jest uiszczana na poziomie organizacji, a nie użytkownika.

Na portalu dostępne są następujące narzędzia:

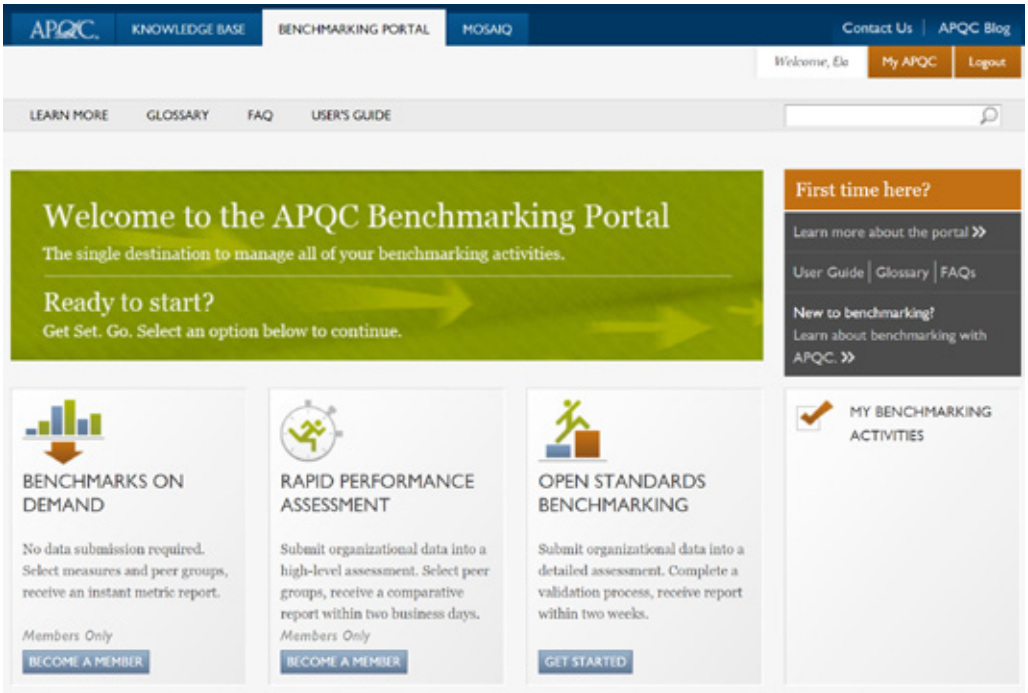
- Usługa „benchmark na żądanie" (ang. *Benchmarks on Demand*) – dostęp do wartości najwyższej, najniższej i mediany określonych wskaźników. Ta usługa jest dostępna natychmiast na żądanie, nie wymaga przesyłania własnych danych, a wskaźniki ograniczone są do tych wyliczonych na podstawie wybranej grupy podmiotów.
- Szybka ocena wydajności (ang. *Rapid Perfomance Assessment*) – raport generowany na podstawie danych opisujących przedsiębiorstwo, identyfikujący problemy wydajnościowe. Wskazuje również na różnice pomiędzy analizowanym przedsiębiorstwem a liderem w danej branży. Czas wygenerowania raportu to 2 dni robocze.
- Open Standards Benchmarking® – szczegółowy raport wydajności, analizujący przedsiębiorstwo na poziomie pojedynczych procesów biznesowych. Wydajność procesów przedsiębiorstwa jest porównywane z innymi jednostkami z branży pod kątem wybranych wskaźników. Czas przygotowania raportu to 2 tygodnie.

Wszystkie trzy narzędzia są dostępne po założeniu konta na portalu (uruchomienie analiz wymaga konta członkowskiego) (rys. **5.13**). Po wybraniu usługi benchmark na żądanie użytkownik może przeglądać zasoby pod kątem słów kluczowych lub kategorii mierników (kategorie

21 https://www.apqc.org/benchmarking-portal

są zgodne z klasyfikacją PCF opisaną w dalszej część rozdziału). Można również zdefiniować grupę porównawczą – kryteria doboru grupy to branża, region geograficzny oraz wolumen przychodów.

RYSUNEK 5.13. **PORTAL APQC – PANEL UŻYTKOWNIKA**



Źródło: https://www.apqc.org/benchmarking-portal, dostęp 30.01.2019 r.

Narzędzie szybkiej oceny wydajności (ang. *Rapid Perfomance Assessment*) dostarcza raport zawierający wartości KPI[22] dla wybranej kategorii procesów organizacji. Formularz oceny organizacji składa się z 15-25 pytań, na które użytkownik musi odpowiedzieć, a końcowy raport zawiera 10-20 miar. Pierwszym krokiem jest wybór kategorii procesów (rys. **5.14**). Następnie należy wypełnić formularz dostarczający informacji o ocenianej organizacji – wartości te są zestawiane z wartościami z grupy porównawczej.

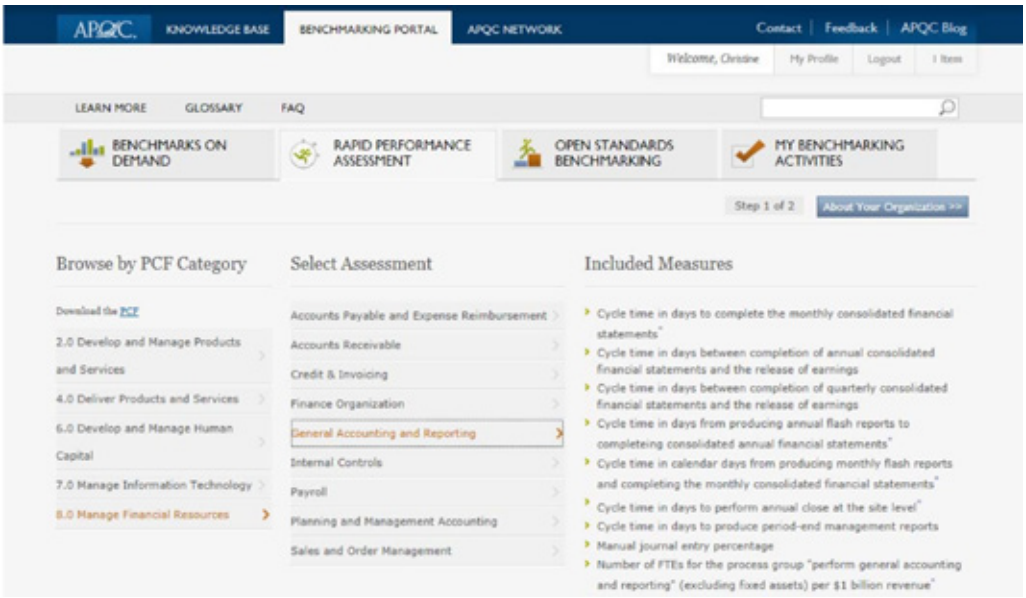
Narzędzie Open Standards Benchmarking® (rys. **5.15**) jako podstawę klasyfikacji mierników przyjmuje PCF. Ten typ oceny jest o wiele bardziej rozbudowany niż szybka ocena wydajności, a formularz oceny organizacji jest bardziej rozbudowany. Wyniki przedstawiają ocenę

22 ang. *Key Performance Indicators* – kluczowe wskaźniki efektywności

poszczególnych procesów organizacji wraz z odchyleniem efektywności ich wykonania od wyników grupy porównawczej. Portal APQC dostarcza też bazy otwartych standardów benchmarkowych i wskaźników wydajności, które pozwalają organizacji przygotować się do analizy. The Open Standards Benchmarking Database zawiera dane z obszarów rachunkowości, zarządzania kapitałem ludzkim, IT, rozwoju produktu, sprzedaży i marketingu oraz zarządzania łańcuchem dostaw.

RYSUNEK 5.14. **PORTAL APQC – NARZĘDZIE SZYBKIEJ OCENY WYDAJNOŚCI**

Źródło: https://www.apqc.org/benchmarking-portal, dostęp 30.01.2019 r.



APQC dostarcza również narzędzia klasyfikacji procesów biznesowych (*Process Classification Framework – PCF*[23]). Jest to taksonomia procesów biznesowych (rys. **5.16**), która dzięki przypisaniu określonym procesom metryk, pochodzących z oferowanych na portalu narzędzi benchmarkingu, pozwala organizacjom śledzić efektywność wykonania procesów i porównać ją z innymi podmiotami w branży. PCF jest podstawą do przeprowadzenia Open Standards Benchmarking®. Framework jest dostępny w wersji uniwersalnej oraz w dziewiętnastu wersjach branżowych[24]. PCF jest aktywnie rozwijany od wczesnych lat 1990. i jest dojrzałym szeroko stosowanym narzędziem oceny wydajności organizacji.

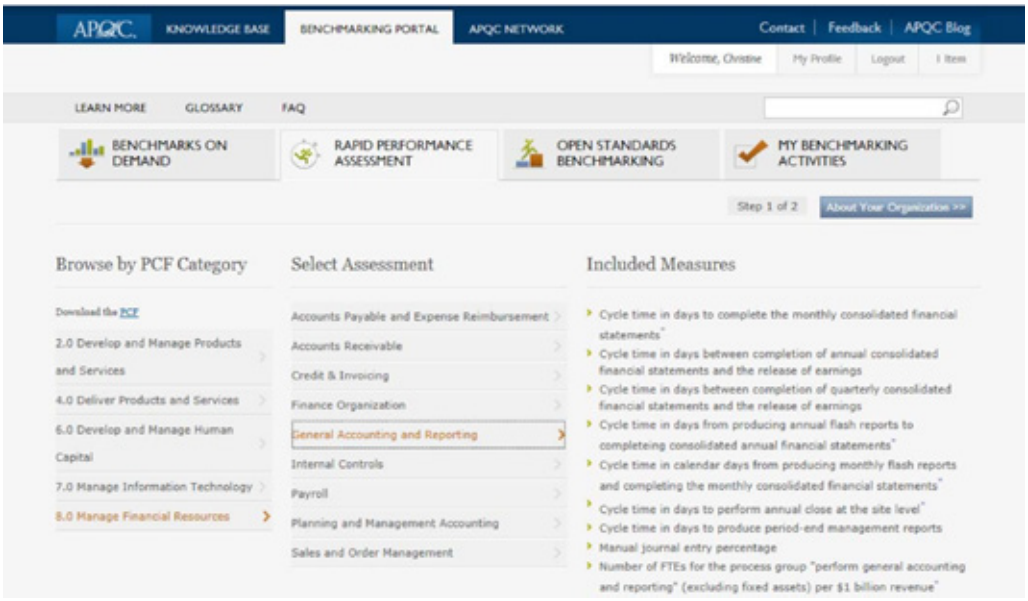
23 https://www.apqc.org/pcf
24 https://www.apqc.org/industry-specificprocess-classification-frameworks

PCF opisuje procesy na pięciu poziomach szczegółowości (rys. **5.17**):

Poziom 1: Kategorie – określają ogólne obszary funkcjonalne przedsiębiorstwa, dla których definiuje się procesy biznesowe.
Poziom 2: Grupy procesów – Bardziej szczegółowe obszary, grupujące procesy z tego samego obszaru funkcjonalnego.
Poziom 3: Procesy – faktyczne procesy biznesowe, uwzględniające różne możliwe przebiegi i warianty czynności.
Poziom 4: Aktywności – Kluczowe elementy (czynności) wykonywane w określonym procesie biznesowym.
Poziom 5: Zadania – Uszczegółowienie aktywności.

Na rys. **5.18** przedstawiono fragment taksonomii wraz ze wskazaniem, dla których elementów zdefiniowane są metryki. Są one dostępne przede wszystkim na poziomie kategorii i grup procesów, jednak dla niektórych elementów możliwy jest również pomiar na niższych poziomach.

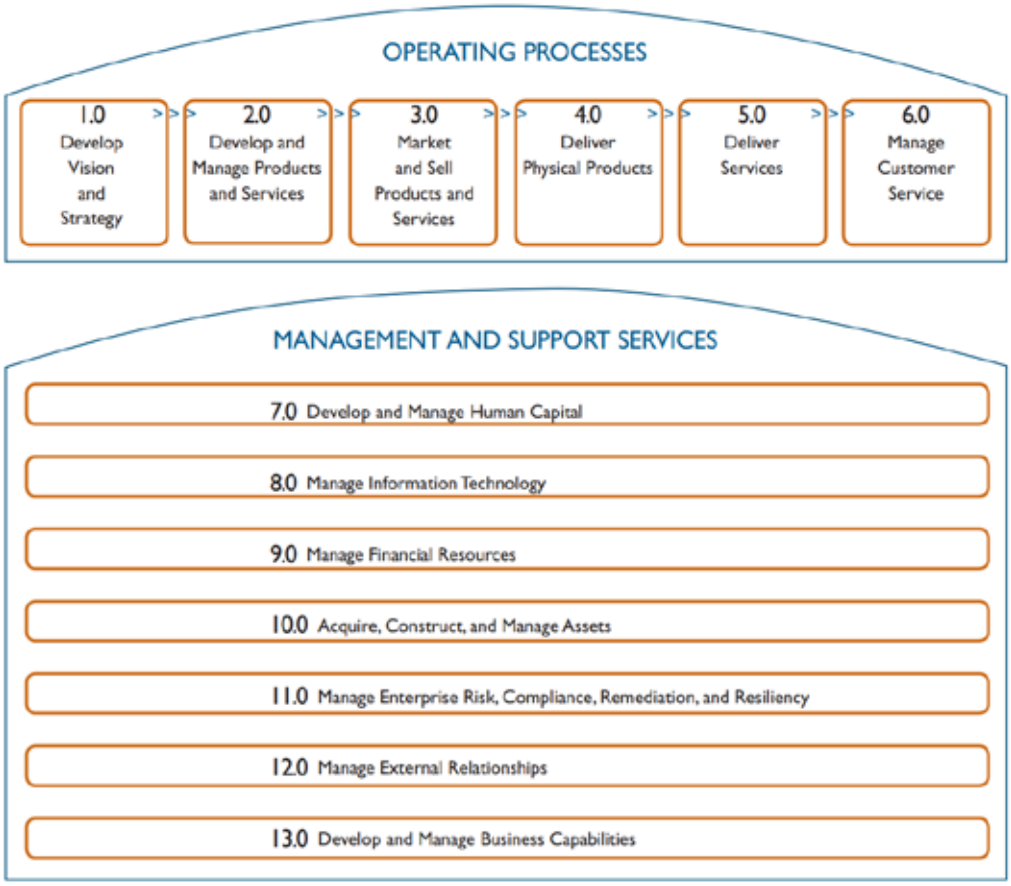
RYSUNEK 5.15. **PORTAL APQC – NARZĘDZIE OPEN STANDARDS BENCHMARKING®**



Źródło: https://www.apqc.org/benchmarking-portal, dostęp 30.01.2019 r.

Portal APQC jest przykładem stworzenia modelu biznesowego wokół danych branżowych. Przykłady zastosowania PCF zostały omówione w sekcji **5.5.4**. Przykłady otwartych źródeł danych, które mogą być wykorzystane na potrzeby budowy usługi benchmarkingu, przedstawiono w rozdziale **3**.

RYSUNEK 5.16. **PCF – KATEGORIE W TAKSONOMII PROCESÓW**



Źródło: https://www.apqc.org/knowledge-base/documents/apqc-processclassification-framework-pcf-cross-industry-pdf-version-721, dostęp 30.01.2019 r.

RYSUNEK 5.17. **PCF – POZIOMY TAKSONOMII**

Źródło: https://www.apqc.org/knowledge-base/documents/apqc-processclassification-framework-pcf-cross-industry-pdf-version-721, dostęp 30.01.2019 r.

RYSUNEK 5.18. **PCF — FRAGMENT TAKSONOMII**

PCF ID	Hierarchy ID	Name	Metrics available?
10007	7.0	**Develop and Manage Human Capital**	Y
17043	7.1	Develop and manage human resources planning, policies, and strategies	Y
20958	7.1.1	Develop human resources strategy	N
10418	7.1.1.1	Identify strategic HR needs	N
10419	7.1.1.2	Define HR and business function roles and accountability	N
10420	7.1.1.3	Determine HR costs	N
10421	7.1.1.4	Establish HR measures	N
10422	7.1.1.5	Communicate HR strategies	N
10432	7.1.1.6	Develop strategy for HR systems/technologies/tools	N
20606	7.1.1.7	Manage employer branding	N
17045	7.1.2	Develop and implement workforce strategy and policies	N
10423	7.1.2.1	Gather skill requirements according to corporate strategy and market environment	N
10424	7.1.2.2	Plan employee resourcing requirements per business	N

Źródło: https://www.apqc.org/knowledge-base/documents/apqc-processclassification-framework-pcf-cross-industry-pdf-version-721, dostęp 30.01.2019 r.

5.4.2. Narzędzia do samooceny

Na rynku istnieje wiele firm oferujących odpłatne usługi benchmarkingu, jednak systematyczne korzystanie z nich może stanowić zbyt wysoki koszt dla niektórych przedsiębiorstw. Rozwiązaniem mogą być darmowe narzędzia i modele samooceny organizacji. Pozwalają one na analizę potencjału organizacji oraz śledzenie kierunku jej rozwoju. Organizacje dostarczające wspomnianych narzędzi niejednokrotnie oferują również usługi konsultingowe, więc możliwa jest okresowa kontrola wybranych wskaźników efektywności i rozwoju wykonana przez ekspertów. Poniżej przedstawiono dwa rozwiązania z tego zakresu.

Model doskonałości EFQM

European Foundation for Quality Management (EFQM) jest organizacją, która opracowała model referencyjny oceny potencjału innowacyjnego przedsiębiorstw – Model Doskonałości EFQM (ang. *The EFQM Excellence Model*). Fundacja EFQM deklaruje, że ich model jest najczęściej używanym narzędziem do ciągłego rozwoju organizacji, możliwym do zastosowania niezależnie od jej wielkości oraz branży, w której działa[25]. Model został opracowany już w roku 1990, a jego najnowsza, zaktualizowana edycja została opublikowana w roku 2013 i uwzględnia dodatkowe obszary związane ze zgodnością działalności z legislacją. Dokumentację modelu można pobrać ze strony EFQM (dostępna również po polsku)[26].

Zgodnie z deklaracją twórców „model doskonałości EFQM jest oparty na zbiorze europejskich wartości, które po raz pierwszy zostały wyrażone w Europejskiej Konwencji Praw Człowieka (1953 r.) oraz w Europejskiej Karcie Społecznej (zrewidowanej w 1996 r.)" [**47**]. Model ma więc wspierać przedsiębiorstwa we wdrażaniu zasad zrównoważonego i społecznie odpowiedzialnego biznesu. Podstawowe cele EFQM to umożliwienie organizacjom określenia miejsca, w jakim się znajdują na drodze do doskonałości: porównanie swoich mocnych i słabych stron z założeniami misji i wizji. Ponadto model zapewnia jednolite słownictwo opisu organizacji i dostarcza ogólne struktury dla procesów zarządzania.

Na EFQM składają się trzy komponenty:

- podstawowe zasady doskonałości,
- model doskonałości EFQM,
- układ logiczny RADAR.

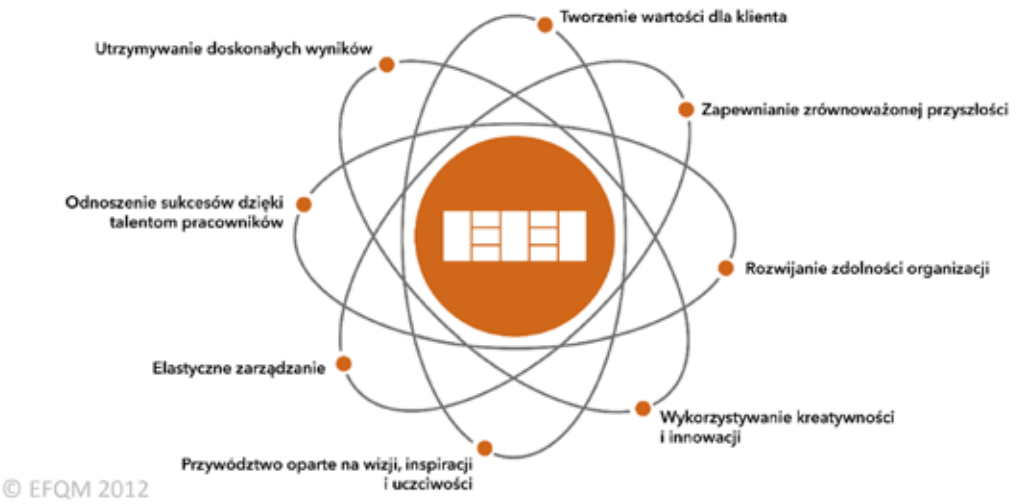
25 http://www.efqm.org/
26 http://www.efqm.org/index.php/efqm-model-2013/download-yourfree-copy/

Model oferowany jest za darmo, ale usługi dodatkowe, takie jak szkolenia czy ocena ekspercka, są płatne.

Zasady doskonałości według modelu EFQM (rys. **5.19**) są następujące:

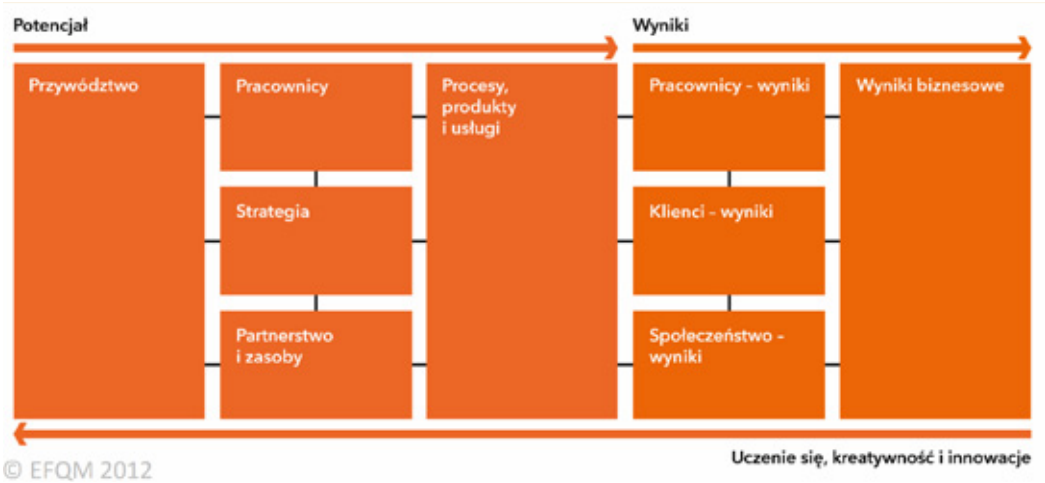
- **Tworzenie wartości dla klienta.** Model EFQM opiera się na założeniu, że doskonałe organizacje są w stanie dostarczyć klientowi wartość, ponieważ rozumieją i przewidują jego potrzeby, równocześnie potrafią skorzystać z pojawiających się w otoczeniu szans. Założenie to stało się podstawą sformułowania „zasad doskonałości", dotyczących kluczowych aspektów zarządzania przedsiębiorstwem, związanych z klientami, zasobami ludzkimi oraz skutecznej analizie dostępnych danych.
- **Zapewnianie zrównoważonej przyszłości.** Model EFQM podkreśla istotność społecznej odpowiedzialności biznesu. Rozwój „doskonałej" organizacji nie może odbywać się kosztem jej otoczenia społecznego, ekologicznego czy ekonomicznego. Zrównoważony rozwój powinien kształtować strategie organizacyjne oraz sposób realizacji procesów biznesowych.
- **Rozwijanie zdolności organizacji.** Ciągły rozwój kompetencji wewnątrz organizacji, jak również w jej otoczeniu jest kluczowy dla osiągnięcia i utrzymania przewagi konkurencyjnej.
- **Wykorzystywanie kreatywności i innowacji.** Wykorzystywanie innowacji jest kluczowe dla rozwoju organizacji. Model EFQM zakłada, że innowacje mogę być pozyskiwane poprzez wykorzystanie zbiorowej wiedzy, a skuteczność ich pozyskiwania zależy od skuteczności w budowaniu sieci uczenia się i współpracy.
- **Przywództwo oparte na wizji, inspiracji i uczciwości.** Liderzy organizacji powinni swoimi działaniami przyczyniać się do budowania kultury organizacyjnej, jak również dawać przykład realizacji misji i wizji organizacji.
- **Elastyczne zarządzanie.** Zdolność do szybkiego i skutecznego reagowania na zmieniające się warunki otoczenia organizacji, wykorzystywanie pojawiających się szans oraz przeciwdziałanie zagrożeniom, jest kluczowe dla powodzenia organizacji na rynku.
- **Odnoszenie sukcesów dzięki talentom pracowników.** Kapitał ludzki jest podstawowym filarem działalności organizacji. Model EFQM wskazuje na konieczność stosowania kultury upełnomocnienia, która pozwala na równoczesne realizowanie celów osobistych i organizacyjnych, a tym samym przyczynia się do wzrostu efektywności organizacji.
- **Utrzymywanie doskonałych wyników.** Celem doskonałych organizacji jest nie tylko osiąganie wybitnych wyników, ale też utrzymanie ich na wysokim poziomie w długim okresie.

RYSUNEK 5.19. **MODEL DOSKONAŁOŚCI EFQM – PODSTAWOWE ZASADY DOSKONAŁOŚCI**



Źródło: [47]

RYSUNEK 5.20. **MODEL DOSKONAŁOŚCI EFQM**



Źródło: [47]

Model Doskonałości składa się z dziewięciu kryteriów, pogrupowanych w dwóch obszarach: „potencjału" oraz „wyników" (rys. **5.20**). „Potencjał" to sposób działania organizacji, jej model biznesowy. „Wyniki" reprezentują osiągnięcia organizacji. Pomiędzy obszarami występuje sprzężenie: *wyniki są efektem potencjału, a potencjał jest doskonalony dzięki wykorzystaniu informacji zwrotnych uzyskanych z wyników* [**47**].

Model EFQM zawiera szczegółową definicję każdego z kryteriów, wraz z przykładami ich realizacji. Kryteria „potencjału" są ściśle powiązane z podstawowymi zasadami doskonałości, co przedstawia rys. **5.21**.

RADAR jest narzędziem samooceny punktowej (rys. **5.22**). Może też służyć ocenie rozwoju organizacji w kolejnych okresach lub do benchmarkingu. Poszczególne elementy podzielone są na atrybuty, które organizacja ocenia punktowo. Model dostarcza też wskazówek, w jaki sposób taka ocena powinna przebiegać.

Na potrzeby analizy potencjału zdefiniowano następujące elementy:

- podejście (atrybuty: pewne, zintegrowane),
- wdrożenie (atrybuty: pełne, ustrukturyzowane),
- ocena i doskonalenie (atrybuty: pomiar, uczenie się i kreatywność, doskonalenie i innowacja).

Na potrzeby analizy wyników zdefiniowano następujące elementy:

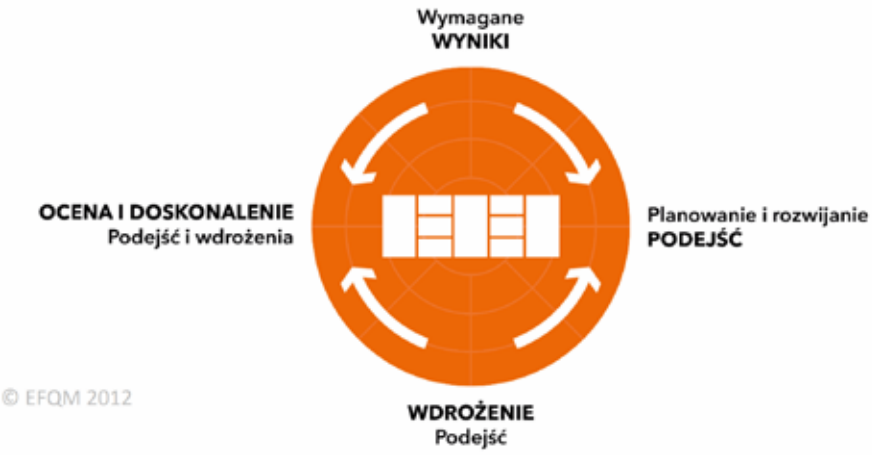
- adekwatność i użyteczność (atrybuty: zakres i adekwatność, wiarygodność, segmentacja),
- poziom (atrybuty: trendy, mierzalne cele, porównania, pewność).

RYSUNEK 5.21. **MODEL DOSKONAŁOŚCI EFQM – POWIĄZANIE KRYTERIÓW POTENCJAŁU I ZASAD DOSKONAŁOŚCI**

Kryterium	1. Przywództwo					2. Strategia				3. Pracownicy					4. Partnerstwo i zasoby					5. Procesy, produkty i usługi				
Podkryterium	a	b	c	d	e	a	b	c	d	a	b	c	d	e	a	b	c	d	e	a	b	c	d	e
Tworzenie wartości dla klienta																				■	■	■	■	■
Zapewnianie zrównoważonej przyszłości	■		■		■			■							■	■				■				
Rozwijanie zdolności organizacji			■				■		■				■		■								■	
Wykorzystywanie kreatywności i innowacji					■				■			■							■					
Przywództwo oparte na wizji, inspiracji i uczciwości	■		■	■	■																			
Elastyczne zarządzanie					■	■			■	■								■		■				
Odnoszenie sukcesów dzięki talentom pracowników											■	■	■	■										
Utrzymywanie doskonałych wyników		■	■			■		■	■										■					

Źródło: [47]

RYSUNEK 5.22. **MODEL DOSKONAŁOŚCI EFQM – UKŁAD LOGICZNY RADAR**



Źródło: [47]

Innovation Portal

Otwarte i darmowe narzędzia do samooceny przedsiębiorstwa na potrzeby benchmarku lub innych decyzji zarządczych są również dostępne na portalach branżowych. Jednym z przykładów jest Innovation Portal[27], który oferuje zasoby wspomagające nauczanie, uczenie się i przeprowadzenie w praktyce procesów zarządzania innowacjami.

Celem portalu jest dostarczenie narzędzi kompleksowo wspierających wszystkie obszary obsługi innowacji: procesy zarządzania innowacjami, analizę kontekstu organizacyjnego, wyszukiwanie i wybór innowacji, implementacja i procesy uczenia się. Aktualnie dostępne są narzędzia z trzech grup: wspierające opracowywanie strategii innowacji dla organizacji[28], wspierające implementację innowacji[29] oraz wspierające zarządzanie procesami budowania wiedzy i opracowywania innowacji[30].

5.5. Analiza przypadków

Wiele organizacji, w szczególności liderów w swoich branżach, korzysta z idei otwartych innowacji implementując ją według własnych potrzeb. Poniżej przedstawiono kilka przykładów takich aktywności. Mają one przede wszystkim formę konkursu innowacyjności lub hackatonu – na podstawie analizy adaptacji idei otwartych innowacji na rynku wydaje się, że ten sposób motywowania uczestników (tj. element rywalizacji oraz nagrody), jest najchętniej stosowany.

5.5.1. Innowacje dla firm

Netflix

Netflix w latach 2006-2009 prowadził konkurs[31] na opracowanie najlepszego algorytmu *collaborative filtering*, który na podstawie ocen wystawionych przez użytkownika, umożliwiałby przewidywanie ocen, jakie wystawi kolejnym filmom.

Zbiór danych uczących obejmował ponad 100 mln ocen (rekordów) wystawionych przez 480 tys użytkowników dla blisko 18 tys filmów. Nagrody były przyznawane za ulepszenie algorytmu Cinematch (algorytm rekomendacji Netflixa) w stosunku do jego efektywności w roku poprzednim. W roku 2009 Netflix przyznał nagrodę główną za opracowanie algorytmu dającego aż o 10% lepszy wynik niż Cinematch.

27 http://www.innovation-portal.info/toolkits/benchmarking/
28 http://www.innovationportal.info/toolkits/innovation-fitness-test/?sort=task
29 http://www.innovationportal.info/toolkits/project-management/
30 http://www.innovation-portal.info/resources/reflection-andaudit-frameworks/
31 https://netflixprize.com/

Konkurs nie był kontynuowany w roku 2010 z powodu problemów prawnych – pojawiły się zarzuty, że użytkownicy portalu mogli zostać zidentyfikowani na podstawie udostępnionego zbioru danych. Sprawa zakończyła się ugodą, ale Netflix nie zdecydował się na kontynuowanie konkursu.

NASA

NASA's Space App Challenge[32] to ogólnoświatowy hackaton, w którym zespoły z całego świata miały za zadanie stworzyć aplikację (lub model) związaną z jednym z problemów konkursowych (rys. **5.23**). Hackatony, w przeciwieństwie do konkursów, mają bardzo krótki czas trwania – w tym przypadku było to 48 godzin. Hackaton organizowany jest co roku od kwietnia 2012 r. i cieszy się coraz większą popularnością (w 2017 roku w wydarzeniu wzięło udział ponad 25 tys. uczestników, z 69 krajów, przesyłając ponad 2 tys. projektów).

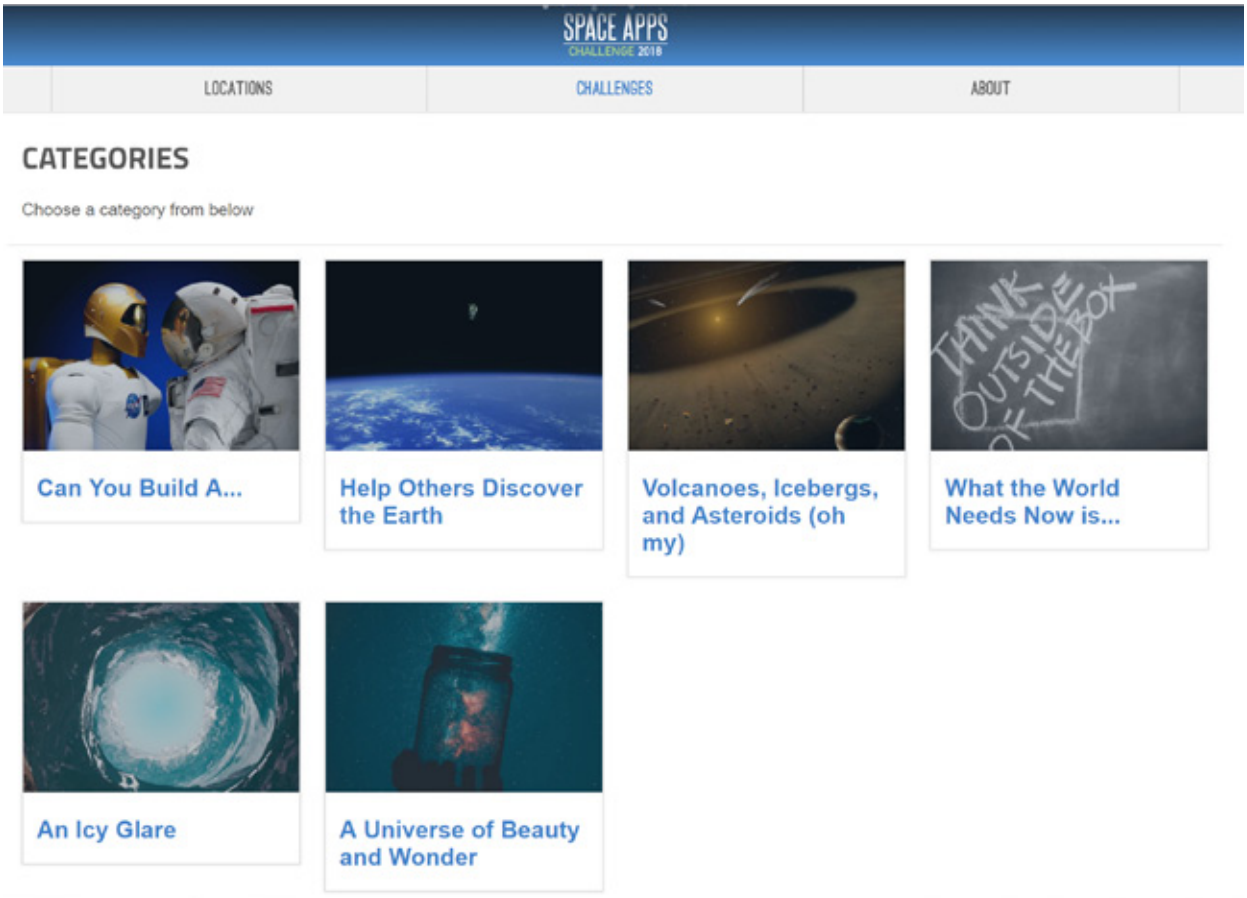
Uczestnicy hackatonu mogli korzystać z publicznie dostępnych danych, w szczególności udostępnianych przez NASA. W każdej edycji organizatorzy przygotowują kilka zadań (tematów przewodnich) dla uczestników. Zawierają one krótki zarys problemu (choć nie tak szczegółowy, jak w przypadku zadań w konkursach innowacyjności) oraz przykładowe zbiory danych, które mogą być wykorzystane w rozwiązaniu (rys. **5.24**). Zbiory danych nie są narzucone uczestnikom, jednak charakter zadań sprzyjał wykorzystaniu otwartych danych dostarczanych przez NASA.

Projekty zgłaszane na hackaton w kolejnych latach były na tak wysokim poziomie, że w roku 2015 NASA utworzyło Space Apps Project Accelerator Toolkit – były to zasoby przeznaczone dla lokalnych organizatorów hackatonów, pozwalające na efektywne budowanie społeczności wokół wydarzenia, w celu przekształcania najlepszych pomysłów w stale rozwijające się innowacje.

Zwycięzcy wybierani są w kilku kategoriach: Best Mission Concept, Best Use of Hardware, Best Use of Data, Most Inspiring, Galactic Impact oraz People's Choice. Taki podział pozwala wyróżnić innowacyjne rozwiązania w wybranym obszarze (np. analizy danych), nawet jeżeli w innych aspektach wymagają jeszcze rozwinięcia.

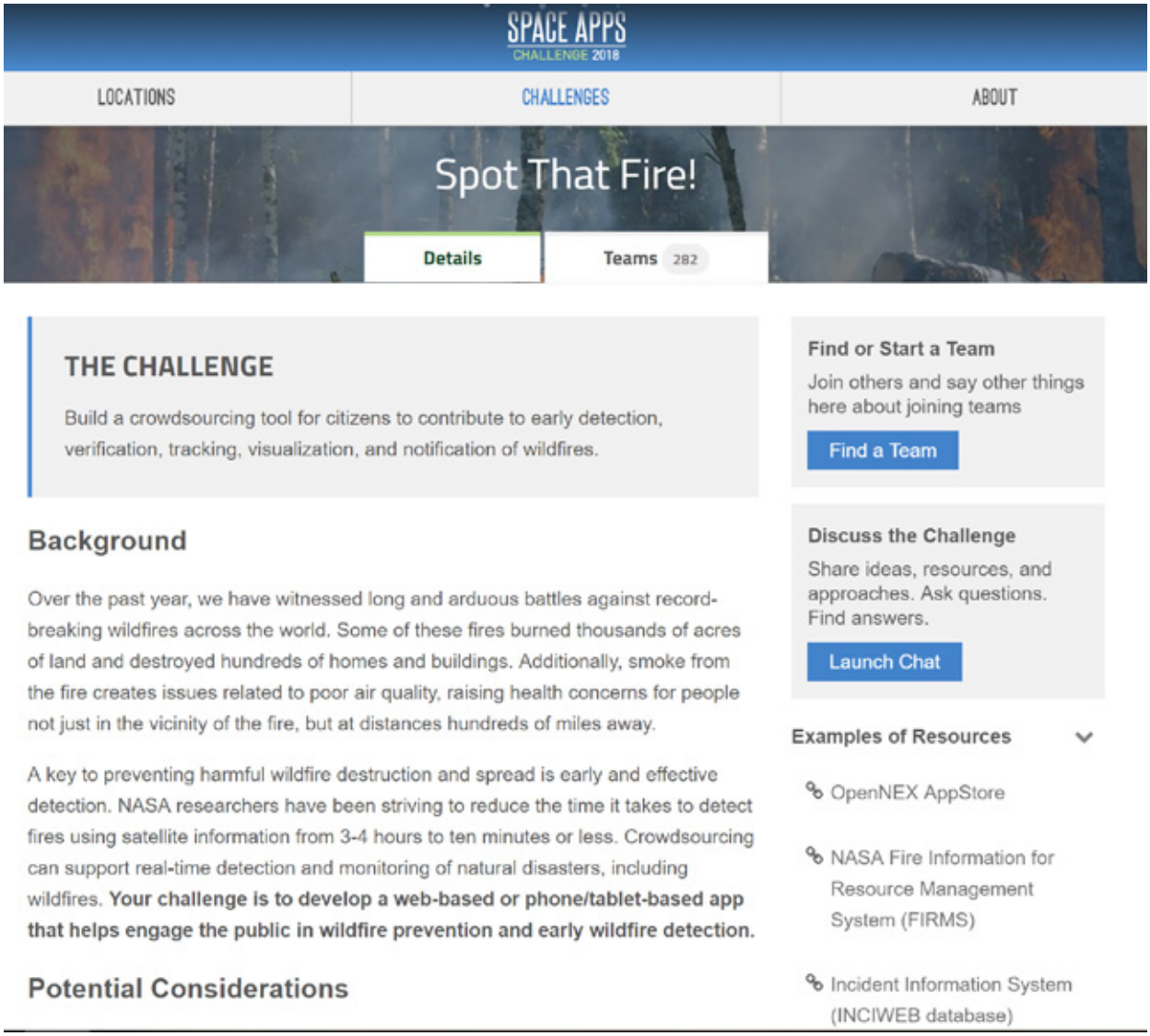
32 https://2018.spaceappschallenge.org/about/

RYSUNEK 5.23. **NASA SPACE APPS HACKATON 2018 – KATEGORIE ZADAŃ KONKURSOWYCH**



Źródło: https://2018.spaceappschallenge.org/challenges/, dostęp 30.01.2019 r.

RYSUNEK 5.24. **NASA SPACE APPS HACKATON 2018 – PRZYKŁADOWE ZADANIE**



Źródło: https://2018.spaceappschallenge.org/challenges/, dostęp 30.01.2019 r.

Konkursy na platformie Kaggle

Konkursy na platformie Kaggle są organizowane przez firmy komercyjne oraz organizacje non-profit czy projekty badawcze. Poza zdefiniowaniem problemu organizatorzy definiują format danych wynikowych oraz udostępniają dane wejściowe (zazwyczaj podzielone na zbiór uczący i testowy).

Przykładowe konkursy innowacyjności na platformie Kaggle:

- Konkurs banku Santander[33] na opracowanie modelu szacującego wartość transakcji potencjalnego klienta. Rozwiązanie miało posłużyć wprowadzeniu personalizowanych usług i poprawie jakości obsługi oraz efektywności pozyskiwania klientów. Uczestnicy otrzymali dwa zbiory zanonimizowanych danych (uczący i testowy). Wyznaczono pięć nagród pieniężnych (odpowiednio: 25/15/10/5/5 tys. USD). Czas trwania konkursu został określony na niecałe dwa miesiące.

  W konkursie wzięły udział 4484 zespoły (5002 uczestników), zgłaszając w sumie 54.898 rozwiązań. Rysunek **5.25** przedstawia listę rankingową po zamknięciu konkursu. Wart uwagi jest fakt, że większość zespołów biorących udział w konkursie jest jednoosobowych, jednak analizując ich profile widać, że często są to reprezentanci firm z branży data science.

- Projekt badawczy The Human Protein Atlas zorganizował konkurs[34] związany z rozpoznawaniem obrazów mikroskopowych. Opracowana metoda miała być wykorzystana w prowadzonych badaniach.

  Sponsorem konkursu były firmy Leica Microsystems oraz NVIDIA – ufundowały cztery nagrody (odpowiednio 14/10/8/5 tys. USD). Konkurs trwał 2 miesiące. Dane wejściowemi ałypostaćobrazów(wpełnejrozdzielnościorazskompresowane)oraz zawierały listę etykiet, którymi należało opisać obrazy. W konkursie wzięły udział 2172 zespoły (2717 uczestników), zgłaszając 55213 rozwiązań.

- Konkurs Google[35], w którym zadanie polegało na opracowaniu algorytmu rozpoznawania odręcznych rysunków pochodzących z ich aplikacji Quick, Draw! Celem było zbudowanie lepszego modelu, niż ten, który był aktualnie wykorzystywany przez aplikację. W konkursie można było zdobyć trzy nagrody: 12/8/5 tys. USD. Trwał on 2 miesiące. Swoje rozwiązania

33 https://www.kaggle.com/c/santandervalue-prediction-challenge
34 https://www.kaggle.com/c/human-protein-atlas-imageclassification
35 https://www.kaggle.com/c/quickdraw-doodle-recognition

zgłosiło 1316 zespołów (1584 uczestników). Zgłoszonych rozwiązań było 21.407. Organizatorzy nałożyli ograniczenia na liczbę zgłoszonych rozwiązań – można było przesłać ich maksylamnie 5 dziennie, a tylko dwa mogły zostać przekazane przez zespół do finalnej oceny.

Udostępniony zbiór danych miał postać plików tekstowych (CSV), zawierających współrzędne wektorowe, z których można było wygenerować grafiki.

- Konkurs organizowany przez Radiological Society of North America[36]. Zadanie konkursowe polegało na analizie obrazów RTG w celu wykrycia anomalii wskazujących na zapalenie płuc. Motywacją do ogłoszenia konkursu była bardzo wysoka zapadalność na pneumonię w Stanach Zjednocznych (w 2015 roku: 500 tys. zachorowań, 50 tys. przypadków śmiertelnych). Równocześnie diagnozowanie tej choroby wymaga eksperckiej analizy zdjęcia RTG, dokładnego wywiadu oraz badań laboratoryjnych. Analiza zdjęcia RTG jest trudna, ponieważ wiele innych schorzeń daje podobny obraz. Ponadto wiele czynników (np. pozycja pacjenta) może wpływać na wynik badania i utrudniać diagnozę.

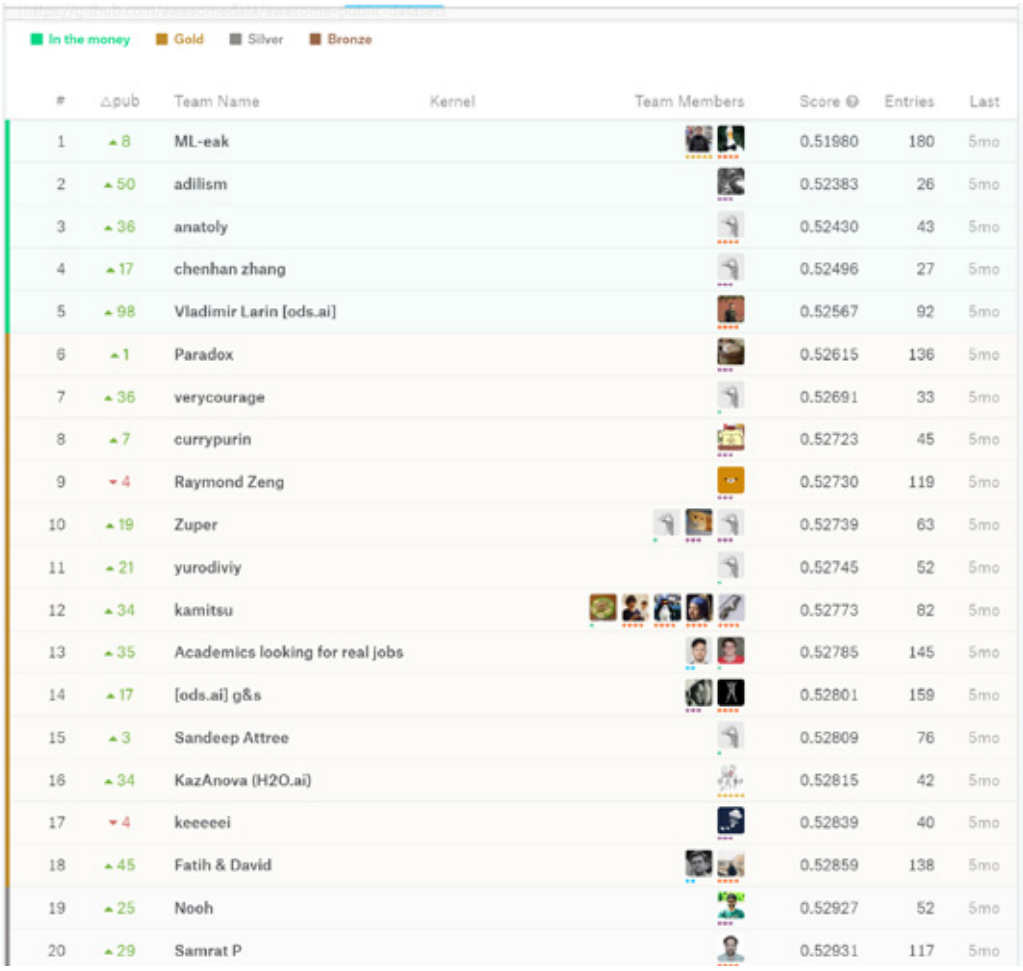
  Konkurs trwał ok. 1,5 miesiąca, wzięło w nim udział 346 zespołów zgłaszając 2020 rozwiązań. Nagrody przyznano dla dziesięciu najlepszych zespołów (kolejne miejsca odpowiednio 12/7/4 tys. USD, miejsca 4-10 po 1 tys. USD).

- Konkurs firmy Airbus[37] na przetwarzanie obrazów satelitarnych w celu szybkiego wykrywania położenia statków. Trudnością w przetwarzaniu obrazów satelitarnych są zakłócenia, np. chmury. Ocenie podlegała nie tylko dokładność wyników, ale też szybkość przetwarzania. Airbus planował wykorzystać rozwiązanie w oferowanej przez siebie usłudze monitorowania ruchu morskiego.

  W konkursie wzięły udział 883 zespoły (1853 uczestników), zgłaszając 12.499 rozwiązań. Wyznaczono nagrody pieniężne dla trzech najlepszych zespołów: 25/15/5 tys. USD oraz nagrodę specjalną dla najszybszego algorytmu: 15 tys. USD. Konkurs trwał 4 miesiące.

36 https: //www.kaggle.com/c/rsna-pneumonia-detection-challenge
37 https://www.kaggle.com/c/airbus-ship-detection

RYSUNEK 5.25. **PLATFORMA KAGGLE – PRZYKŁAD KONKURSU INNOWACYJNOŚCI**



In the money | Gold | Silver | Bronze

#	△pub	Team Name	Kernel	Team Members	Score	Entries	Last
1	▲8	ML-eak			0.51980	180	5mo
2	▲50	adilism			0.52383	26	5mo
3	▲36	anatoly			0.52430	43	5mo
4	▲17	chenhan zhang			0.52496	27	5mo
5	▲98	Vladimir Larin [ods.ai]			0.52567	92	5mo
6	▲1	Paradox			0.52615	136	5mo
7	▲36	verycourage			0.52691	33	5mo
8	▲7	currypurin			0.52723	45	5mo
9	▼4	Raymond Zeng			0.52730	119	5mo
10	▲19	Zuper			0.52739	63	5mo
11	▲21	yurodiviy			0.52745	52	5mo
12	▲34	kamitsu			0.52773	82	5mo
13	▲35	Academics looking for real jobs			0.52785	145	5mo
14	▲17	[ods.ai] g&s			0.52801	159	5mo
15	▲3	Sandeep Attree			0.52809	76	5mo
16	▲34	KazAnova (H2O.ai)			0.52815	42	5mo
17	▼4	keeeeei			0.52839	40	5mo
18	▲45	Fatih & David			0.52859	138	5mo
19	▲25	Nooh			0.52927	52	5mo
20	▲29	Samrat P			0.52931	117	5mo

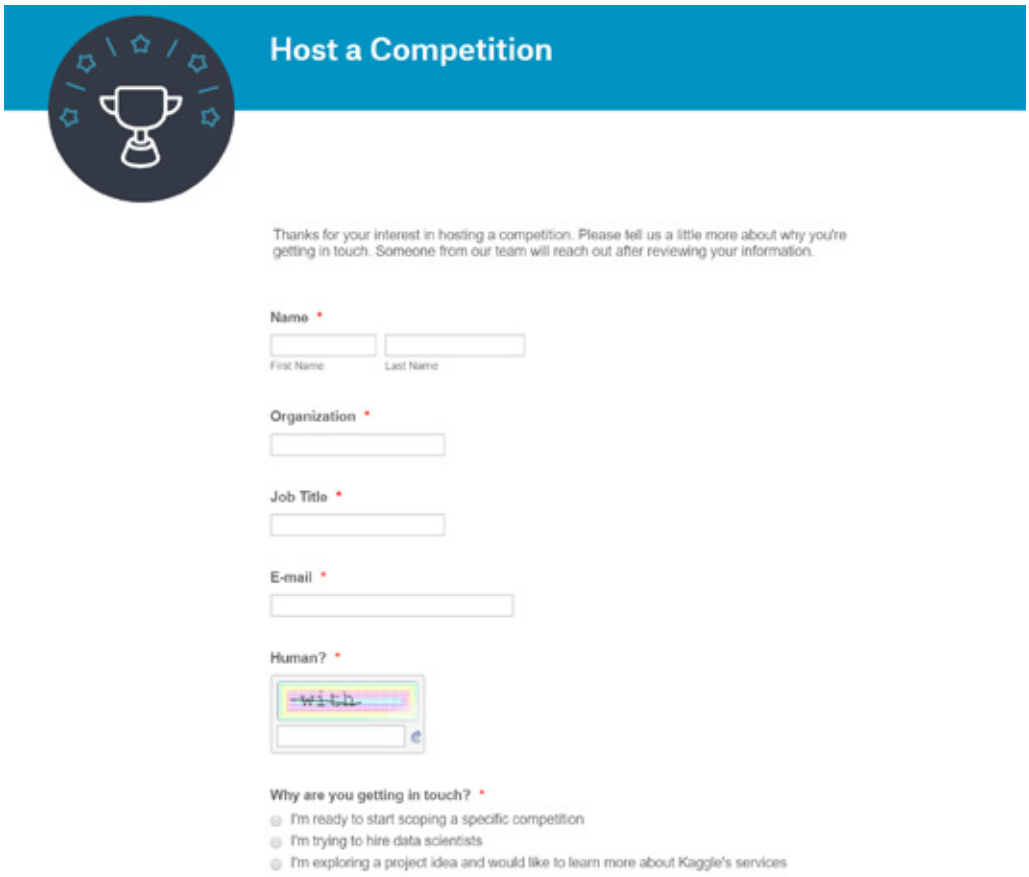
Źródło: https://www.kaggle.com/competitions, dostęp 30.01.2019 r.

W celu zorganizowania konkursu, konieczne jest indywidualne zgłoszenie do Kaggle (rysunek **5.26**). Ustalenie szczegółów konkursu trwa zwykle od 4 do 6 tygodni. Czas trwania konkursu jest ustalany indywidualnie, ale zwykle mieści się w przedziale od 6 do 24 tygodni. Wyniki powinny zostać ogłoszone w ciągu miesiąca od zakończenia konkursu.

Kaggle wymaga nagród pieniężnych w konkursach z kategorii *featured* w wysokości min. 25 tys. USD, ale w zamian wszyscy nagrodzeni muszą dostarczyć organizatorowi (firmie) kod zwycięskiego rozwiązania (zwykle w języku R lub Python) wraz z dokumentacją oraz udzielić licencji,

umożliwiającej komercyjne wykorzystanie rozwiązania. Ponadto organizator ponosi koszt opłaty na rzecz Kaggle. Całkowity budżet potrzebny na organizację konkursu jest szacowany na 85-200 tys. USD (w tym nagrody).

RYSUNEK 5.26. **PLATFORMA KAGGLE – FORMULARZ ZGŁOSZENIA ORGANIZACJI KONKURSU**

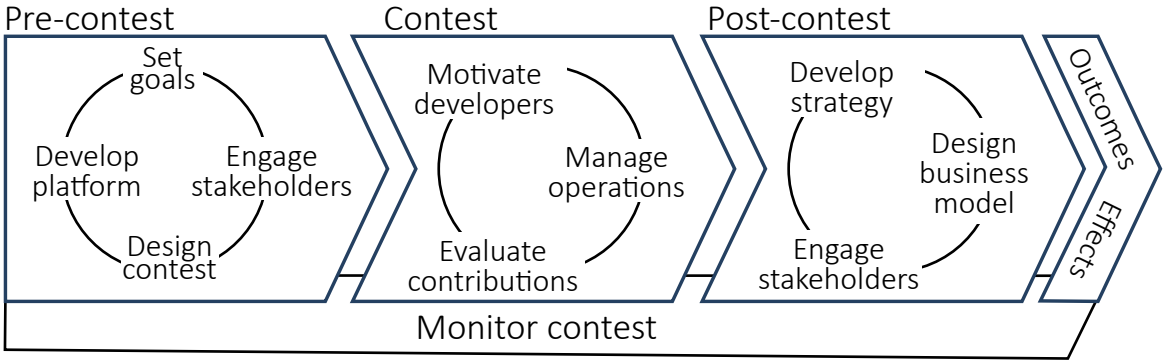


Źródło: https://www.kaggle.com/hosting-inquiry?Host_Business, dostęp 30.01.2019 r.

5.5.2. Proces organizacji konkursu innowacyjności

Hjalmarsson, Juell-Skielse i Johannesson [**74**] przedstawiają ustrukturyzowane podejście do organizowania udanych konkursów innowacyjności (rys. **5.27**), wskazując równocześnie na krytyczne elementy każdego z etapów.

RYSUNEK 5.27. **PROCES ORGANIZACJI KONKURSU INNOWACYJNOŚCI**



Źródło: [74]

Autorzy podejścia podkreślają, że stanowi ono uzupełnienie tradycyjnych procesów opracowywania produktów i usług, powinno więc być z nimi ściśle zintegrowane. Sukces całego procesu jest uwarunkowany współpracą pomiędzy zewnętrznymi deweloperami (uczestnikami konkursu) a organizatorem i beneficjentem, którzy powinni mieć ścisłą kontrolę nad przebiegiem konkursu. Proponowane podejście składa się z trzech głównych faz oraz jedenastu bardziej szczegółowych aktywności.

W fazie przedkonkursowej najważniejszym i kluczowym krokiem jest zdefiniowanie celów konkursu, tj. zadania konkursowego, pożądanych rezultatów i ich formatu, jak również zasad oceny. Większość platform konkursowych oferuje wsparcie dla organizatorów i beneficjentów już na tym etapie – głównymi instrumentami są konsultacje i warsztaty z ekspertami. Etap definiowania celów może trwać nawet kilka tygodni. Bezpośrednio z celów wynika również podział na role w konkursie oraz ich zakresy odpowiedzialności. Kolejne dwa kroki są często realizowane przez platformy konkursowe: jest to zaangażowanie podmiotów powiązanych z konkursem (w tym przypisanie ich do ról) oraz uruchomienie platformy konkursowej. Ostatnim krokiem w tym etapie jest zaprojektowanie konkursu: określenie jego ram czasowych, wymagań formalnych, mediów, przy pomocy których będzie reklamowany, itp.

Druga faza procesu obejmuje faktyczny czas jego trwania. Obejmuje kroki związane z pozyskiwaniem i motywowaniem uczestników. Metody wykorzystane w tym kroku są zależne od wymagań stawianych uczestnikom – im większe wymagania wstępne, tym mniejsza, ale lepiej dopasowana grupa potencjalnych uczestników. Uczestnicy mogą być motywowani np. nagrody lub dostarczanie ciągle aktualizowanej listy rankingowej. Ponadto w tym etapie wykonywane są czynności związane z bieżącym zarządzaniem konkursem. Etap kończy ocena rozwiązań i wyłonienie zwycięzcy.

Z punktu widzenia wdrażania otwartych innowacji najistotniejszy jest etap pokonkursowy, który skupia się na czynnościach związanych z wdrożeniem wypracowanych wyników. Czynności związane z opracowaniem strategii mogą mieć większe lub mniejsze znaczenie w zależności od specyfiki konkursu. Organizator i beneficjent powinni wypracować odpowiednią formę współpracy z uczestnikami (a w szczególności ze zwycięzcami) po zakończeniu konkursu. Często strategia wykorzystania rezultatów jest częściowo określana już na etapie projektowania konkursu, jednak jej finalna wersja może zostać sformułowana dopiero po uzyskaniu określonych rezultatów. Krok znoszenia barier dotyczy rozstrzygania różnych problemów związanych z przejęciem praw do rozwiązania konkursowego, określenie procedur jego rozwoju, integracji z rozwiązaniami beneficjenta, likwidacji błędów i luk. Ostatni obszar to opracowanie modelu biznesowego, pozwalającego monetyzować opracowaną innowację.

Przez cały czas trwania procesu wykonywane są również zadania związane z monitorowaniem konkursu. Mają one na celu wczesną identyfikację potencjalnych problemów i wprowadzenie działań zaradczych.

Istotne jest, aby od początku procesu planowania konkursu innowacyjności brać pod uwagę wszystkie etapy, nie koncentrując się wyłącznie na środkowym. Etap przedkonkursowy jest kluczowy dla precyzyjnego zdefiniowania celów i zakresu innowacji, tak aby odpowiadały one potrzebom biznesowym organizacji. Natomiast etap pokonkursowy pozwala na efektywne wykorzystanie rezultatów.

5.5.3. Innowacje sektorze farmaceutycznym

Open Innovation są wykorzystywane nie tylko na indywidualne potrzeby organizacji, ale również w badaniach zakrojonych na szeroką skalę, np. w przemyśle farmaceutycznym. Brahmachari [**27**] opisuje otwarty program odkrywania leków (*an open source drug discovery programme – OSDD*[38]). Program prowadzony jest w Indiach. Wizją OSDD było stworzenie globalnej platformy, która pozwoliłaby współpracować niezależnym ekspertom (naukowcom, lekarzom, klinikom, szpitalom, itp.) w rozwiązywaniu złożonych problemów związanych z opracowywaniem

38 http://www.osdd.net/about-us

nowych leków i kuracji na choroby, które są nieco lekceważone przez przemysł farmaceutyczny, np. gruźlicę, malarię lub leiszmaniozę. Twórcy platformy (CSIR India Consortium) zainspirowali się otwartoźródłowym modelem tworzenia oprogramowania (ang. *open source*), który z powodzeniem jest też stosowany w biotechnologii (np. do sekwencjonowania genomu). Celem programu jest stworzenie przystępnej cenowo opieki zdrowotnej, w szczególności w zakresie wymienionych wcześniej chorób.

Struktura programu OSDD przedstawiona jest na rys. **5.28**. SysBorg jest wewnętrzną platformą programu, umożliwiającą uczestnikom wymianę informacji, wyników prac, dyskusje i współpracę przez Internet. Dzięki platformie uczestnicy programu mogą w sposób ciągły generować nowe pomysły i alternatywne rozwiązania postawionych problemów, co napędza proces opracowywania innowacji. Najlepsze pomysły są kierowane do dokładniejszych badań laboratoryjnych, a stamtąd do prób klinicznych, których powodzenie może się zakończyć wprowadzeniem produktu na rynek.

Uczestnikami programu są również przedstawiciele biznesu (firm skupionych wokół CSIR). W programie bierze udział 7900 uczestników (studentów, przedstawicieli nauki i biznesu oraz przedstawicieli korporacji) z ponad 130 krajów.

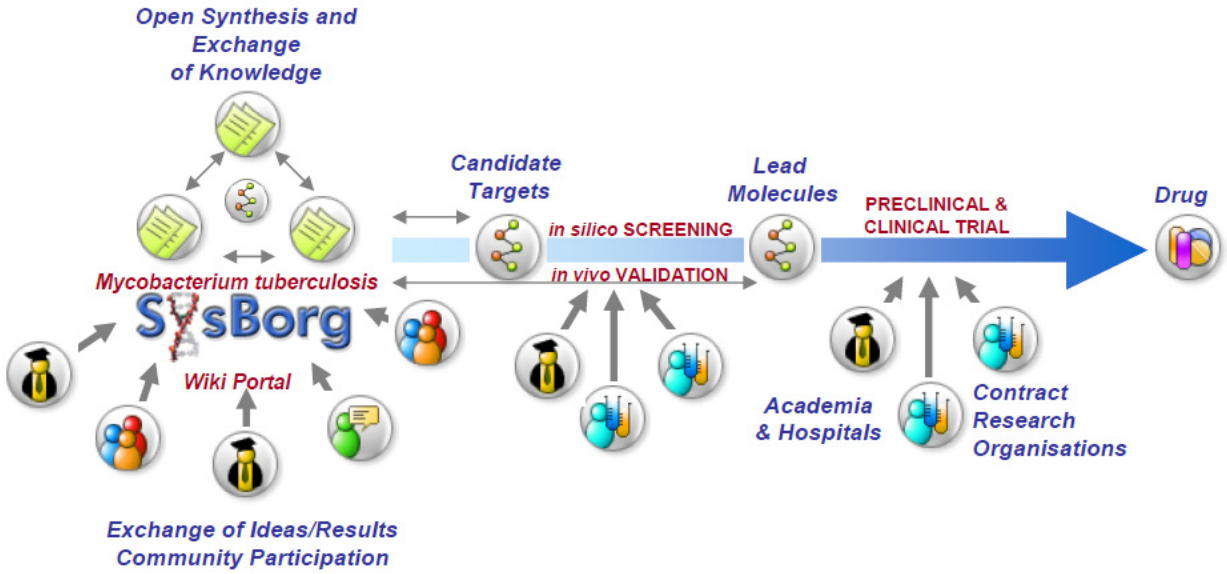
OSDD wykorzystuje idee otwartych innowacji, łącząc je z crowdsourcingiem i modelem open source, aby zmienić dominujący w branży farmaceutycznej model biznesowy. Program koncentruje się na lekach na te choroby, które dla dużych firm farmaceutycznych nie stanowią interesującego obszaru badań, ponieważ występują w biedniejszych regionach, przez co zwrot z inwestycji na poniesione badania może być bardzo niski. Wyniki badań OSDD są udostępnione, więc każda firma może z nich skorzystać i zacząć produkować leki, bez potrzeby ponoszenia dodatkowych kosztów na ich opracowanie. Zwiększa to konkurencyjność w branży i pozwala na obniżenie cen.

Rysunek 5.29. przedstawia proces opracowywania leku w programie OSDD. Na wczesnych etapach badań, polegających na generowaniu innowacyjnych pomysłów, dyskusjach, prototypowaniu, prace wykonywane są przez społeczność. W jej skład wchodzą zarówno bardziej doświadczeni badacze, jak i studenci czy młodzi naukowcy. Natomiast w drugim etapie badań, polegającym na szczegółowych badaniach i testach, prace są zorganizowane w sformalizowane procesy, a do ich wykonania wykorzystuje się doświadczone organizacje badawcze z całego świata. Ostatni etap, próby kliniczne, są finansowane ze środków publicznych.

Wyniki przeprowadzonych badań dostępne są na stronie internetowej projektu[39]. Co więcej, na stronie dostępne jest też specjalistyczne repozytorium danych biologicznych i chemicznych.

39 http://www.osdd.net/research-development

RYSUNEK 5.28. **STRUKTURA PROGRAMU OSDD**



Źródło: http://www.osdd.net/about-us/how-osdd-works, dostęp 30.01.2019 r.

Przykład OSDD pokazuje zastosowanie otwartych innowacji, skierowanej na rozwój niekomercyjny. Znaczące jest, że program nie oferuje nagród finansowych (jak robią to konkursy innowacyjności), a mimo to angażuje dużą grupę uczestników i pozwala na opracowywanie i wprowadzanie na rynek innowacyjnych rozwiązań.

5.5.4. Benchmarking

Rozwiązania benchmarkingowe APQC są szeroko stosowane przez organizacje na całym świecie.

Intel

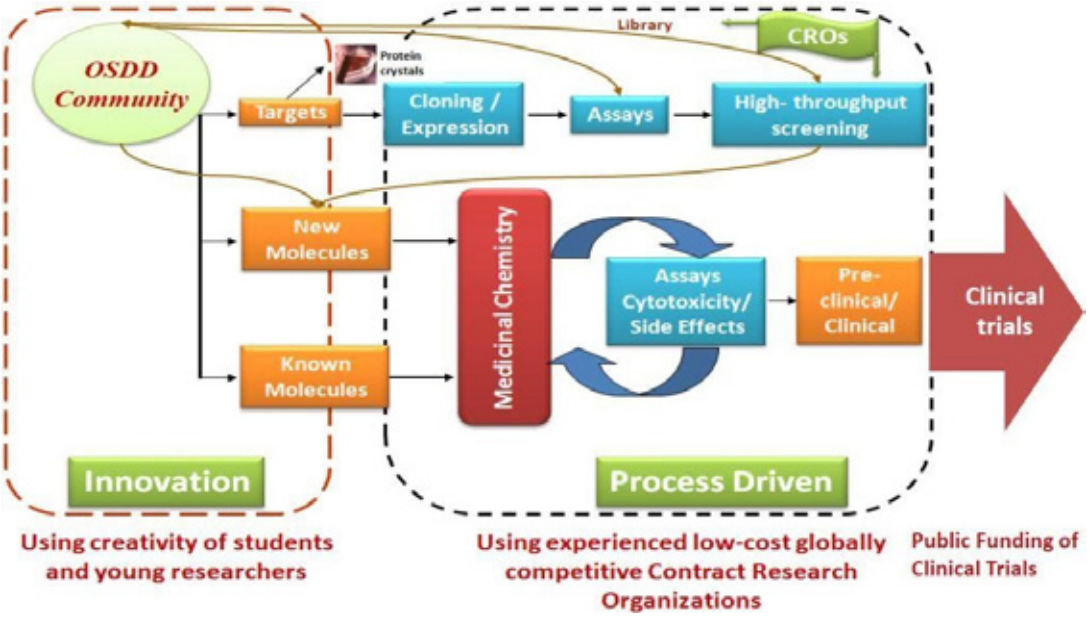
Intel zastosował rozwiązania APQC na potrzeby zarządzania jakością informacji[40]. Celem projektu było przeniesienie metod zarządzania jakością wyrobów Intela na spójny system zarządzania jakością informacji. Firma zdecydowała się na analizę benchmarkingową rozwiązań z tego zakresu stosowanych przez konkurencję, ponieważ nie udało się zidentyfikować gotowego do wykorzystania rozwiązania.

40 https://www.apqc.org/knowledge-base/documents/managinginformation-quality-intel

Na zlecenie Intela zespół APQC przeprowadził badanie online oraz bezpośrednie wywiady na grupie 67 przedstawicieli różnych organizacji, z których 25 należało do grupy porównawczej Intela (organizacje pozausługowe, o rocznym przychodzie ponad 1 miliard USD). Badanie trwało od maja do września 2012 roku. Wyniki badania wykazały, że postawiony problem nie dotyczy wyłącznie Intela, ale jest charakterystyczny dla branży. Ponadto, choć każda z badanych organizacji korzystała z własnego podejścia do wspierania strategii jakości informacji, to udało się wykazać, że największy sukces odnoszą modele hybrydowe, opierające się na współpracy między obszarami IT, biznesowymi oraz innymi związanymi z jakością.

Na podstawie wyników badań benchmarketingowych Intel opracował własną metodologię zarządzania jakością informacji, pozycjonując się jako lider w branży, pozwalającą przede wszystkim na przeprowadzenie szybkiej oceny przez pojedynczego pracownika. Proponowane rozwiązania wprowadza proste metryki pomiaru jakości, których wynik zawiera się w przedziale od 0 do 10. Przykładowo, jeżeli zbiór danych jest oceniony na 7 lub mniej, to nie można na jego podstawie przeprowadzić rzetelnych analiz Business Intelligence. Punktowy system oceny jakości informacji pozwolił też Intelowi na identyfikację słabych punktów architektury systemu informacyjnego oraz wyizolować problemy związane z danymi.

RYSUNEK 5.29. **PROCES OPRACOWYWANIA LEKU W PROGRAMIE OSDD**



Źródło: http://www.osdd.net/research-development/discovery-developmentstatergy, dostęp 30.01.2019 r.

Elevations Credit Union

Elevations Credit Union to firma działająca od 1952 roku, która w latach 2009-2012 zanotowała wzrost wartości aktywów na poziomie 37% (z 937 mln USD do 1,3 mld USD). Firma wskazuje na związek między dynamicznym wzrostem a podejmowanymi od 2008 roku działaniami, mającymi na celu wprowadzenie procesowego podejścia do zarządzania organizacją[41].

Wdrażanie podejścia procesowego zostało zaplanowane na okres 5 lat. W każdym roku zdefiniowano cele do osiągnięcia oraz kluczowe aktywności:

Rok 1 Budowanie wstępnej mapy procesów – w oparciu o klasyfikację procesów PCF przygotowano mapę procesów na poziomie grup procesów, a następnie stopniowo uszczegóławiano je na kolejne poziomy zgodnie z frameworkiem.

Rok 2 Zapewnienie spójności w organizacji – stworzenie narzędzi do monitorowania procesów oraz szkolenia z zarządzania procesami biznesowymi.

Rok 3 Dopełnienie procesu w celu wsparcia adaptacji rozwiązania – podejście procesowe na tym etapie było już częścią kultury organizacyjnej i możliwy był monitoring, analiza i benchmarking procesów. Prace skupiały się na rozwinięciu metod zarządzania procesami biznesowymi i uspójnieniu przepływu informacji.

Rok 4 Wykorzystanie transparentności doskonalenia procesów do angażowania organizacji – firma Elevations upubliczniła rezultaty udoskonalania wybranych procesów, co motywowało pracowników do dalszej pracy nad usprawnianiem procesów oraz rozwijania samej dokumentacji, która te procesy opisuje.

Rok 5 Utrzymanie rozwiązania – w ostatnim etapie rozwijano elementy związane z zarządzaniem wiedzą (w tym opracowane polityki zarządzania procesami, sformalizowano reguły biznesowe i inne elementy procesowe).

Elevations stworzyło modele 261 procesów (ze szczegółowością na poziomie pojedynczych czynności wykonywanych przez pracowników najniższych szczebli), każdy z nich powiązany z PCF. PCF dla branży bankowej służy jako przyjęty przez organizację język opisu procesów. Ponadto uzyskano skrócenie czasu wykonania procesu udzielania pożyczek: z 10-13 dni do 4-5 dla pożyczek na zakup pojazdów oraz z 30 dni do 14-15 dni dla pożyczek na zakup nieruchomości. Średnia liczba pożyczek konsumenckich w miesiącu wzrosła z 483 do 659.

41 https://www.apqc.org/knowledgebase/documents/using-apqcs-process-classificationframework-pcf-elevations-credit-union-ca

5.6. Podsumowanie

W rozdziale omówiono inicjatywy związane z ideą open innovation. Proces opracowywania otwartych innowacji jest wykonywany w dużej mierze poza organizacją – zamiast koncentrować się na wynikach własnych prac badawczych, firmy przekazują część czynności do wykonania przez społeczność. Wiąże się to jednak z koniecznością rozwiązania kwestii udostępniania danych. Ze względu na ochronę prywatności klientów oraz tajemnice handlowe, przedsiębiorstwa mogą sceptycznie podchodzić do otwartych innowacji. Przykłady konkursów innowacyjności, jakie są z sukcesem organizowane w różnych branżach, pokazują jednak, że możliwe jest włączenie społeczności do prac badawczych, z równoczesnym poszanowaniem prywatności. Firmy stosują w tym celu kilka różnych rozwiązań:

- udostępnienie danych zanonimizowanych,
- udostępnienie ograniczonego zbioru danych (bez danych wrażliwych),
- wskazanie otwartych źródeł danych, na których należy opracować prototyp rozwiązania,
- udostępnienie danych pod warunkiem przestrzegania przez użytkowników restrykcyjnych licencji.

W przypadku konkursów innowacyjności dane są najczęściej udostępniane w postaci gotowych repozytoriów, opisane metadanymi, zwykle podzielone na zbiory testowe i uczące (a organizator często wykorzystuje jeszcze inny zbiór do oceny zgłoszonych rozwiązań). Ich format jest różny, gdyż zwykle dostosowuje się go do problemu badawczego, ale dane mają zazwyczaj stosunkowo wysoką jakość (często są to dane z wewnętrznych zasobów firmy).

Jak opisane w podrozdziale **5.2**, firmy mogą brać udział w konkursach innowacyjności w różnych rolach: jako organizator, beneficjent, dostawca zasobów lub uczestnik. Każda z tych ról wiąże się z innymi obowiązkami i korzyściami.

Warte uwagi są pojawiające się inicjatywy związane z otwartymi algorytmami – najbardziej zaawansowaną platformą tego typu jest OPAL (patrz podrozdział **5.3**). OPAL umożliwia wykonanie określonych algorytmów na danych prywatnych, w taki sposób, aby zwracane były wyłącznie zagregowane wyniki.

Ostatni obszar omówiony w rozdziale to narzędzia do analizy porównawczej. Jest to obszar, w którym wciąż brakuje otwartych źródeł danych, umożliwiających porównywanie się firm między sobą. Jest to z jednej strony spowodowane niechęcią do ujawnienia danych stanowiących o przewadze konkurencyjnej, a z drugiej wysokim stopniem komercjalizacji usług benchmarkingu. Jednakże takie inicjatywy jak EFQM (patrz sekcja **5.4.2**) pokazują, że ta branża również zaczyna się skłaniać ku otwartym rozwiązaniom.

6 Open science – nauka dla społeczności

6.1. Wprowadzenie

Koncepcja otwartej nauki (*open science*) jest często łączona z koncepcją otwartych danych, choć ta pierwsza pojawiała się już w XVII wieku. Na początku była ona rozumiana jako powszechny dostęp do wiedzy. Dziś znacznie częściej kojarzona jest z dostępem do danych wykorzystywanych lub wytworzonych podczas prowadzenia badań. Faktycznie, pojęcie otwartej nauki zakłada, „że priorytetowy jest nie tylko swobodny dostęp do ostatecznego rezultatu prac badawczych, czyli zapewnienie otwartego dostępu do artykułów naukowych, ale również stosowanie otwartych modeli w innych obszarach pracy naukowej, czyli np. udostępnianie surowych danych”[1].

Popularność koncepcji otwartej nauki wynika przede wszystkim z rozwoju internetu i oferowanych przez niego łatwych możliwości upowszechniania nauki. Nowoczesne technologie niewątpliwie ułatwiają przetwarzania, przechowywanie i dystrybucję treści. Otwarty dostęp to nie tylko dostępność w Internecie, ale przede wszystkim swoboda w zakresie wykorzystania surowych danych: tworzenie własnych kopii, przenoszenie między komputerami, wyszukiwanie, drukowanie, rozpowszechnianie, łączenie z innymi zbiorami. Dostęp powinien być możliwy bez finansowych, prawnych czy technicznych ograniczeń.

Wydawnictwo Springer Nature przeprowadziło jedno z większych badań dotyczących Open Science, w którym odpowiedzi na ankietę udzieliło ok. 7700 naukowców. Wyniki zostały podsumowane w raporcie „Practical challenges for researchers in data sharing” [**11**, **165**]. Główny wniosek jest taki, że sami naukowcy są zainteresowani tym, aby dane wytworzone w ich badaniach były dostępne dla innych. Ujawnia się to w dążenie społeczności badaczy, ujawnionym w politykach, strategiach, grupach roboczych, do tego, aby dane wykorzystane w badaniach były znajdowalne, dostępne, łatwe do powiązania z innymi oraz gotowe do ponownego wykorzystania. Cechy te są określane angielskim akronimem **FAIR** – **F**indable, **A**ccessible, **I**nteroperable and **R**eusable. Open Access, czyli swobodny dostęp do publikacji naukowych może przyspieszyć tempo tworzenia nowych odkryć naukowych: unikanie duplikacji badań. Jest to coraz popularniejszy model, chętnie również wspierany przez Komisję Europejską. Z kolei otwieranie danych w koncepcji open science pozwala na zwiększenie weryfikowalności wyników badań.

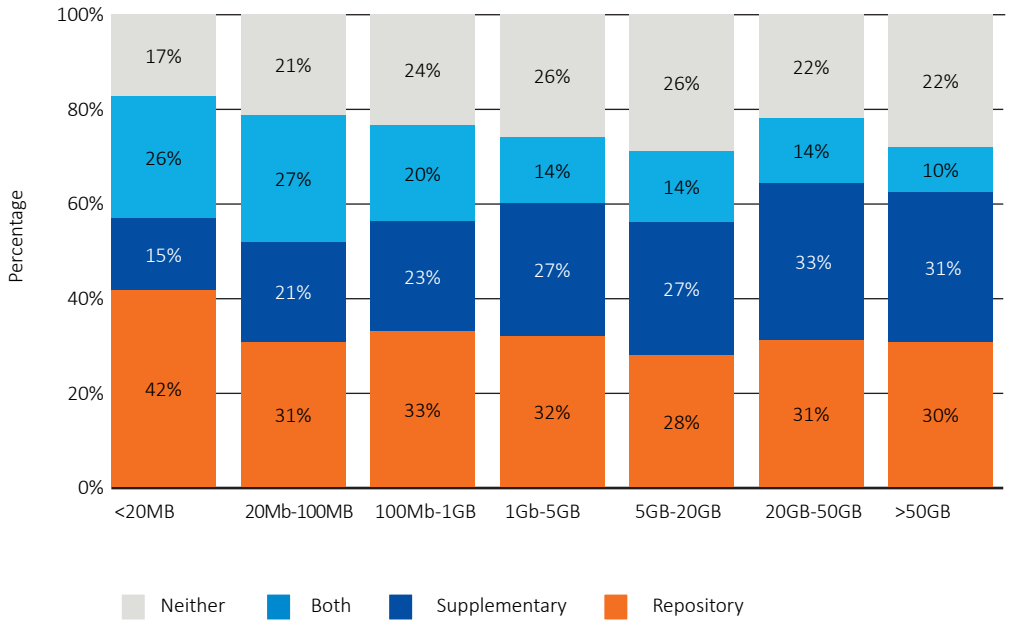
Jeśli chodzi o charakterystykę liczbową, to zaobserwowano wzrost zainteresowania open science. 65% respondentów wskazało, że co do zasady zgłaszają pliki z danymi jako informację dodatkową do ich publikacji lub umieszczają dane w dedykowanym repozytorium. Oceniając w skali 1-10 istotność możliwości znalezienia danych z własnych publikacji naukowcy jako najpopularniejszą wskazali odpowiedź 10 (25% odpowiedzi). Łącznie 76% naukowców bardzo wysoko ocenia tę potrzebę. Ze współdzieleniem danych ze społecznością wiążą się określone

1 https: //otwartanauka.pl/wprowadzenie/czym-jest-otwarta-nauka

wyzwania. Wskazywane były przede wszystkim [**165**]: organizowanie danych w prezentowalnym i użytecznym formacie (46% odpowiedzi), niepewność co do praw autorskich i licencji – 37%, trudność w identyfikacji właściwe repozytorium do publikacji – 33%, brak czasu, aby przygotować dane do publikacji – 26%, koszty współdzielenia danych – 19%.

Szczegóły dotyczące skłonności do dzielenia się danymi w zależności od wielkości zbioru przedstawiona na rysunku.

RYSUNEK 6.1. **ZACHOWANIE DOTYCZĄCE PUBLIKACJI DANYCH NAUKOWYCH W ZALEŻNOŚCI OD WIELKOŚCI ZBIORU**



Źródło: [165]

Skłonność do publikacji zbiorów danych zależy od reprezentowanej dziedziny nauki. Przekłada się to na podaż danych o określonej zawartości. Najwyżej oceniana jest istotność publikacji danych biologii, naukach o ziemi, naukach medycznych oraz fizyce. Naciski na publikację danych badawczych pojawiają się również ze strony wydawców oraz sponsorów badań, którym zależy na uzyskaniu odpowiedniego zwrotu z inwestycji.

6.2. Inicjatywy

W związku z rozwojem polityk na poziomie unijnym, co z kolei wynika z ogromnych funduszy, które są inwestowane w badania, pojawiły się inicjatywy, które te polityki materializują. Dwie flagowe inicjatywy unijne to OpenAIRE oraz EOSC.

6.2.1. OpenAIRE

OpenAIRE[2] to organizacja unijna, której celem jest wspieranie otwartości w komunikacji naukowej. Nazwa jest akronimem dla „Open Access Infrastructure for Research in Europe". Organizacja ta powstała, aby wspomóc naukowców realizujących projekty w programie ramowym Horizon 2020 w raportowaniu publikacji na portalu *EC Participant Portal* oraz ułatwić zachowanie zgodności z *European Commission Open Access Policy and Research Data Pilot*. OpenAIRE osiąga to poprzez agregację wyników badań finansowanych ze środków unijnych z ponad 1000 różnych repozytoriów, a następnie udostępnienie ich na swoim portalu. Statystyki dotyczące liczby różnych zasobów przedstawia rys. **6.2**. Organizacja ta służy również jako punkt konsultacyjny w zagadnieniach otwartego dostępu (Open Access), świadczy usługi analityki badań oraz zarządza płatnościami związanymi z publikacjami o otwartym dostępnie (Article Processing Charges – APCs). Zostali wyznaczeni eksperci z poszczególnych krajów w ramach „National Open Access Desks", łącznie 34 osoby.

RYSUNEK 6.2. **OPENAIRE W STATYSTYKACH**

Źródło: https://provide.openaire.eu/landing

2 https://www.openaire.eu/

Zawartość OpenAIRE jest podzielona na 5 portali tematycznych, w zależności od docelowego odbiory.

- **OpenAIRE.EXPLORE**[3] – docelowym odbiorcą jest badacz. Oferowana jest możliwość eksploracji wyników publikowanych w ramach otwartego dostępu (Open Access). Możliwe jest również podlinkowanie wszystkich własnych wyników badań i stworzenie swojego profilu badacza. Obecnie dostępne jest 25 milionów publikacji, 671 tysięcy zbiorów danych badawczych od 13 tysięcy dostawców oraz 17 instytucji fundujących, realizujących ponad 2,5 miliona projektów. Przeglądarka zbiorów została przedstawiona na rys. **6.3**.
- **OpenAIRE.PROVIDE**[4] – docelowym odbiorcą są dostawy treści. Oferowana jest możliwość lepszej promocji własnych zasobów. Może to być osiągnięte dzięki usłudze walidacji metadanych i sprawdzenia zgodności ze wskazówkami OpenAIRE. Dostawcy mogą zarejestrować własne repozytoria i połączyć je z innymi zasobami, tym samym wzbogacając własne metadane. Istotna jest również możliwość mierzenia oddziaływania własnych treści i odbioru społecznego. Obecnie na tę część składa się 3615 repozytoriów literatury, 2249 repozytoriów danych, 19710 czasopism z otwartych dostępem, 117 agregatorów.
- **OpenAIRE.CONNECT**[5] – docelowym odbiorcą są grupy badawcze. Oferowana jest możliwość pozyskania wiarygodnego partnera w celu współdzielenia, rozpowszechniania i monitorowania publikacji naukowych. Ta część jest na razie w wersji beta.
- **OpenAIRE.MONITOR**[6] – docelowym odbiorcą są kierownicy badań (czy to w projektach, czy instytucjach). Oferowana jest możliwość monitorowania publikacji, zarówno pod kątem ich pojawienia się w pierwotnym źródle (repozytorium), jak też referencji do nich. W tej części nacisk jest wyraźnie na analitykę.
- **OpenAIRE.DEVELOP**[7] – docelowym odbiorcą są deweloperzy. Oferowana jest możliwość programowego dostępu do zawartości zbieranej w ramach OpenAIRE. Przedsiębiorcy są zachęcani do tworzenia własnych produktów w oparciu o dane – do monetyzacji zasobów oferowanych w ramach otwartej nauki. Jako sposób komunikacji wykorzystywany jest OAI-PMH[8] – standardowy protokół do publikacji i ściągania metadanych.

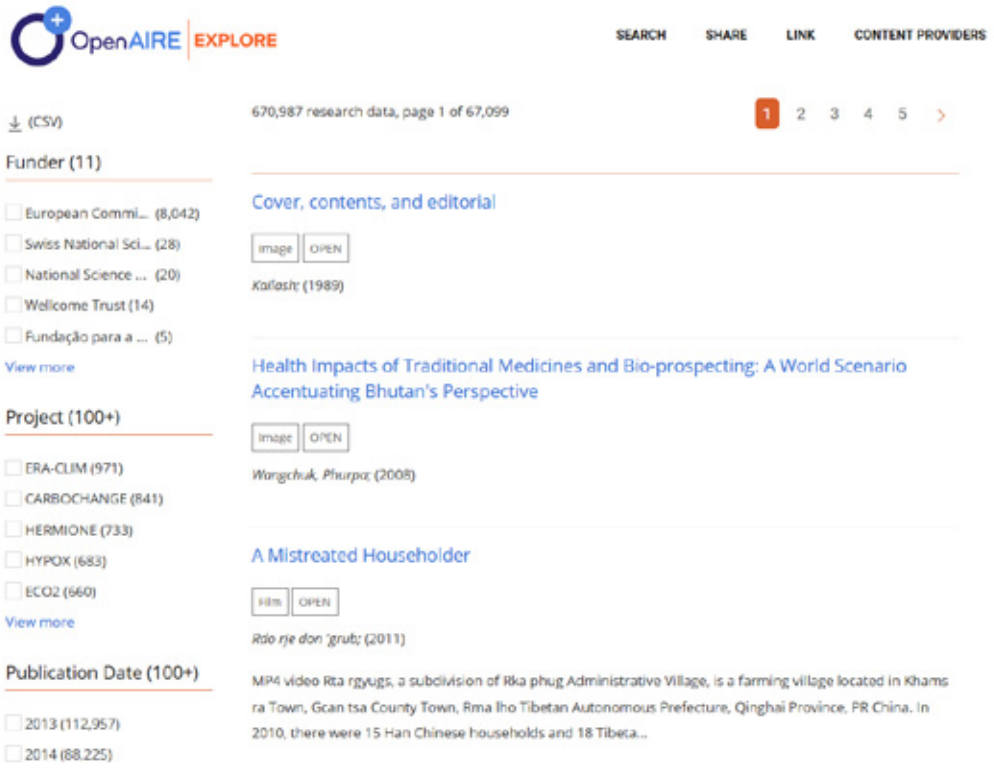
Endpoint to tworzenia zapytań znajduje się pod adresem http://api.openaire.eu/oai_pmh a opis API pod adresem http://develop.openaire.eu/api.html.

3 https://explore.openaire.eu/
4 https://provide.openaire.eu/landing
5 https://connect.openaire.eu/
6 https://monitor.openaire.eu/
7 http://develop.openaire.eu/
8 The Open Archives Initiative Protocol for Metadata Harvesting, http://www.openarchives.org/OAI/openarchivesprotocol.html

Na przykład w celu otrzymania informacji o projekcie LOD2 należy wpisać następujący adres do przeglądarki: http://api.openaire.eu/search/projects?name=lod2. Listę projektów z Polski, które zajmowały się badaniami nad mózgiem zwróci następujące zapytanie: http://api.openaire.eu/search/projects?name= brain&participantCountries=PL. Liczby zbiorów wg typów przedstawiają się następująco: Dataset 537.323, Image 78.276, Sound 8.207, Film 5.913.

Głównym produktem OpenAIRE jest Zenodo, repozytorium pozwalające na publikację wyników badań. Zostało ono opisane poniżej w sekcji **6.5.5**.

RYSUNEK 6.3. **PRZEGLĄDANIE ZBIORÓW W OPENAIRE EXPLORE**



Źródło: https://explore.openaire.eu/search/find/datasets

6.2.2. European Open Science Cloud

European Open Science Cloud (EOSC[9]) jest kolejną inicjatywą unijną. Jest to inicjatywa zmierzająca do opracowania polityk na wysokim poziomie, wspierających przede wszystkim Open Science i Open Innovation. Zajmuje się wspieraniem dobrych praktyk w dostępnie do danych (m.in. FAIR). Została powołana grupa ekspertów, których zadaniem jest wypracowanie roadmapy dla europejskiej chmury otwartej nauki. Roadmapa wyznacza sześć obszarów, w których będą określane szczegółowe działania: a) architektura, b) dane, c) usługi, d) dostęp i interfejsy, e) reguły oraz f) zarządzanie (governance).

W kwietniu 2018 roku OpenAIRE oraz EOSC-hub porozumiały się co do współpracy na poziomie unijnym we wdrażaniu filozofii współdzielenia wyników badań naukowych[10]. Plan współpracy opiera się na trzech filarach: a) integracja usług, b) komunikacja, wsparcie i szkolenia oraz c) zarządzanie i strategia.

6.3. Otwarte repozytoria nauki

Otwarta nauka to przede wszystkim możliwość „pochwalenia" się osiągnięciami przez naukowców. Najczęściej przyjmuje to formę publikacji, a zbiory danych są jedynie interesującym dodatkiem. W niniejszej sekcji dokonujemy krótkiej analizy inicjatyw otwartych repozytoriów w Polsce.

6.3.1. PON – Platforma Otwartej Nauki

W polskiej nauce główną inicjatywą jest Platforma Otwartej Nauki (PON)[11]. Misją PON jest zwiększenie dostępności, widoczności i oddziaływania dorobku polskich naukowców. Zgodnie z informacją na stronie jest to „centrum kompetencji w zakresie otwartych modeli komunikacji naukowej i dystrybucji wiedzy, ośrodek oferujący rozwiązania i narzędzia umożliwiające wdrażanie tych modeli, a także cyfrowa kolekcja polskich publikacji naukowych.". PON zajmuje się zarówno opracowywaniem dobrych praktyk, jak i tworzeniem narzędzi do ich wdrażania.

Rozwija również serwisy, które pozwalają na gromadzenie i udostępnianie informacji naukowej. W swoich serwisach udostępnia różne zasoby: przede wszystkim są to polskie czasopisma naukowe (*Biblioteka Nauki*), książki naukowe (*Otwórz Książkę*) oraz różnego rodzaju prace naukowe (w większości artykuły) deponowane przez samych autorów (Repozytorium CeON). Biblioteka Nauki[12] we współpracy z bazami bibliograficznymi udostępnia polskie czasopisma naukowe. Na koniec stycznia 2019 roku w bazie znajdowało się 278879 artykułów

9 https://ec.europa.eu/research/openscience/index.cfm?pg=openscience-cloud
10 https://www.openaire.eu/eosc-hub-and-openaireadvance-get-together-to-support-open-science-and-theeuropean-open-science-cloud-initiative
11 http://pon.edu.pl/
12 http://yadda.icm.edu.pl/yadda/search/general.action

z 911 czasopism, które były bezpłatnie dostępne dla każdego. Otwórz książkę[13] to z kolei cyfrowa kolekcja publikacji naukowych. Na koniec stycznia 2019 roku 408 autorów udostępniło 544 książki, które można pobrać lub przeczytać w serwisie.

6.3.2. Repozytorium Centrum Otwartej Nauki CEON

Repozytorium Centrum Otwartej Nauki CEON[14] z założenia dedykowane jest zbieraniu publikacji polskich naukowców w otwartym dostępie. W istocie, można znaleźć tutaj wyłącznie publikacje w liczbie ok. 15 tys. Z pobieżnej analizy wynika, że większość artykułów poświęcona jest naukom humanistycznym.

Repozytorium jest zgodne z protokołem pobierania metadanych opracowanym przez OpenArchive Initiative (OAI). Dzięki temu specjalne serwisy, które zajmują się agregacją informacji o zasobach naukowych w formie cyfrowej mogą łatwo odnaleźć odpowiednie treści.

Nazwa repozytorium mogłaby sugerować zbieranie również danych źródłowych, ale to nie ma tutaj miejsca. CEON nie jest przydatny z punktu widzenia polskich przedsiębiorców.

6.3.3. Repozytorium Danych Otwartych

Rzeczywiste zbiory danych można z kolei znaleźć w Repozytorium Danych Otwartych RepOD[15]. Jest ono częścią PON, choć, co ciekawe, brak linku do niego ze strony głównej PON. Repozytorium to zostało zbudowane z wykorzystaniem oprogramowania CKAN[16], stąd też osoby poszukujące dane trafią na znany już im interfejs oraz filozofię deponowania plików.

W repozytorium znajduje się zaledwie 91 zbiorów (datasetów). Zarejestrowanych jest 149 użytkowników. Wydzielone zostały 4 grupy: 1) PAN Brain Imaging and Functional Research of Nervous System, 2) ICM UW, 3) IPPT PAN, 4) Laboratory of Neuroinformatics, Nencki Institute. Łącznie dostępne jest ok 500 GB danych, przy czy istotną część zajmują obrazowania medyczne (dane ze zdjęć rentgenowskich[17]).

Największy zbiór plików (dataset) to wyniki symulacji z największego dostępnego publicznie, fizjologicznie realistycznego, sieciowego wzoru mózgu szczura[18].

13 http://otworzksiazke.pl/
14 https://depot.ceon.pl/
15 https://repod.pon.edu.pl/
16 http://ckan.org/
17 https://repod.pon.edu.pl/dataset/raw-x-ray-diffraction-datafor-medicago-truncatula-omega-amidase-in-ada-buffer
18 https://repod.pon.edu.pl/dataset/thalamocortical-network

6.4. Bazy publikacyjne

Wspomniane w poprzedniej sekcji repozytoria nauki nastawione są na polskie zasoby. Obecna nauka nie zna granic. Dla kompletności obrazu należy mieć również rozeznania w zagranicznych repozytoriach artykułów. Poniżej prezentujemy najistotniejsze inicjatywy, w których można znaleźć interesujące publikacje. Skupiamy się na tych, które oferują otwarty dostęp. Wyłączamy więc zbiory dostępne tylko z bibliotek na zasadzie subskrypcji.

Jako ciekawe inicjatywy w obszarze otwartości nauki należy wskazać dwa portale. Ich główną wartością nie są jednak same dane, a analityka. Altmetric[19] ma odpowiedzieć na podstawowe pytanie: kto mówi o naszych badaniach. W tym celu śledzi określone źródła i znajduje odwołania do prac naukowych. Docelowymi odbiorcami są wydawcy, instytucje badawcze, naukowcy i fundatorzy badań. Plum Analytics[20] to rozwiązanie dedykowane mierzeniu oddziaływania badań naukowych. Firma została założona w 2011 roku, a w 2017 przejął ją Elsevier. Zajmuje się wyliczaniem alternatywnych miar jakości publikacji, tzw. *altmetrics*. Włączane są miary związane z popularnością czy też odbiorem społecznym, np. komentarze na Twitterze, dyskusje na blogach, zakładki w Mendeley, cytowania w Scopus i inne.

6.4.1. Mendeley

Firma Mendeley została założona w 2007 roku przez 3 niemieckich doktorantów. Nazwa została wymyślona dla uczczenia dwóch wielkich badaczy: Mendla oraz Mendelejewa. W 2013 roku firma została przejęta przez wydawnictwo Elsevier, co wywołało ogromne kontrowersje i wzbudziło dyskusję dotyczącę wolnego dostępu oraz open science.

Mendeley to oprogramowanie służące do zarządzania i współdzielenia własnych publikacji oraz odkrywania nowych artykułów i danych. Składa się na nie aplikacja desktopowa oraz część webowa w postaci portalu z elementami sieci społecznościowej, pomiędzy którymi można dokonywać synchronizacji. Aplikacja desktopowa charakteryzuje się przejrzystym interfejsem (zob. rys. **6.4**) i bardzo wysoką użytecznością. Wiele działań można wykonać w sposób intuicyjny. Do najważniejszych ułatwiających pracę badacza funkcji zaliczamy:

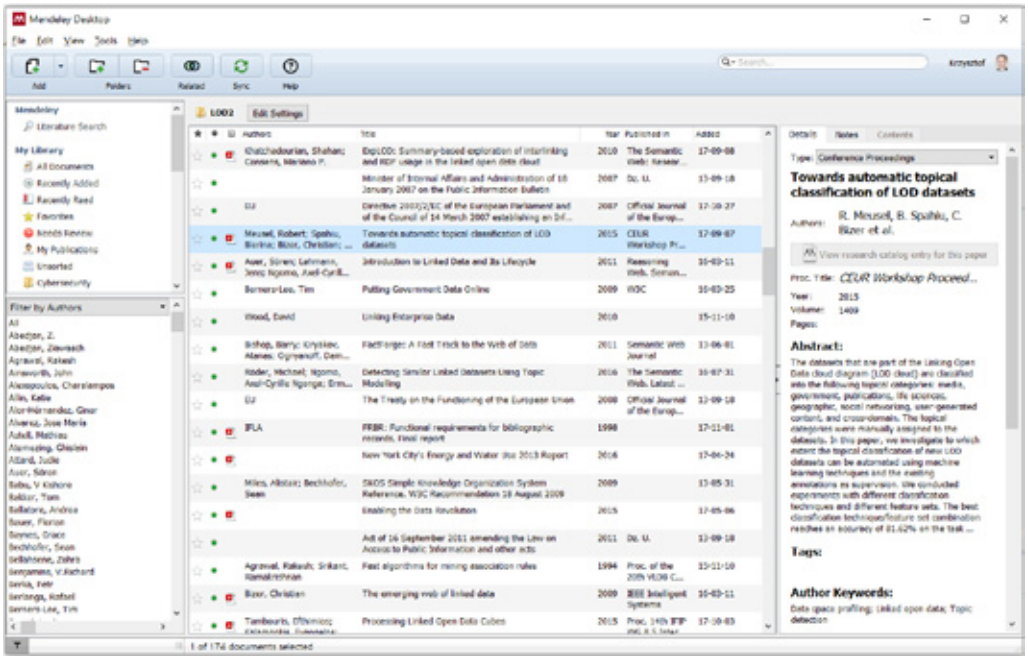
- importowanie plików PDF i tworzenie kolekcji w podziale na kategorie,
- automatyczne tworzenie metadanych publikacji poprzez analizę plików PDF oraz dociąganie informacji udostępnionych przez innych użytkowników,
- plugin do programu Word, który umożliwia szybkie i łatwe cytowanie prac, a także importowanie nowych metadanych bezpośrednio z dokumentów Word,

19 https://www.altmetric.com/
20 https://plu.mx/

- możliwość tworzenia backupu i synchronizacji plików online, przez co zapisane publikacje dostępne są na dowolnym urządzeniu,
- możliwość eksportu do formatu BibTeX – możliwe cytowanie prac w LaTeX-u,
- tworzenie notatek bezpośrednio w plikach PDF,
- pełnotekstowe wyszukiwanie w zaimportowanych plikach PDF.

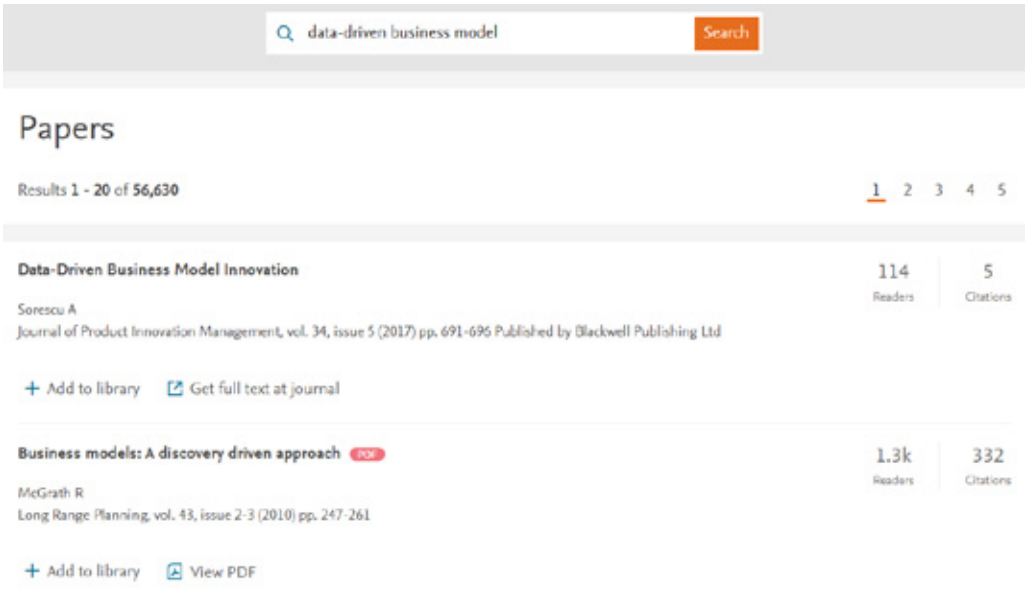
Główna funkcjonalność portalu wiąże się z możliwością wyszukiwania artykułów wg słów kluczowych. Rys. **6.5** przedstawia wyniki wyszukiwania dla frazy 'data-driven business model'. W wyniku otrzymaliśmy ok. 57 tys. artykułów. Przy każdym artykule otrzymujemy informację, ile razy był cytowany oraz ile osób dodało go do swojej biblioteki. Do niektórych artykułów dołączony jest plik PDF, inne są opatrzone tylko linkiem do strony czasopisma.

RYSUNEK 6.4. **INTERFEJS APLIKACJI DESKTOPOWEJ MENDELEY**



Źródło: zrzut ekranu.

RYSUNEK 6.5. **WYNIKI WYSZUKIWANIA ARTYKUŁÓW W PORTALU MENDELEY**

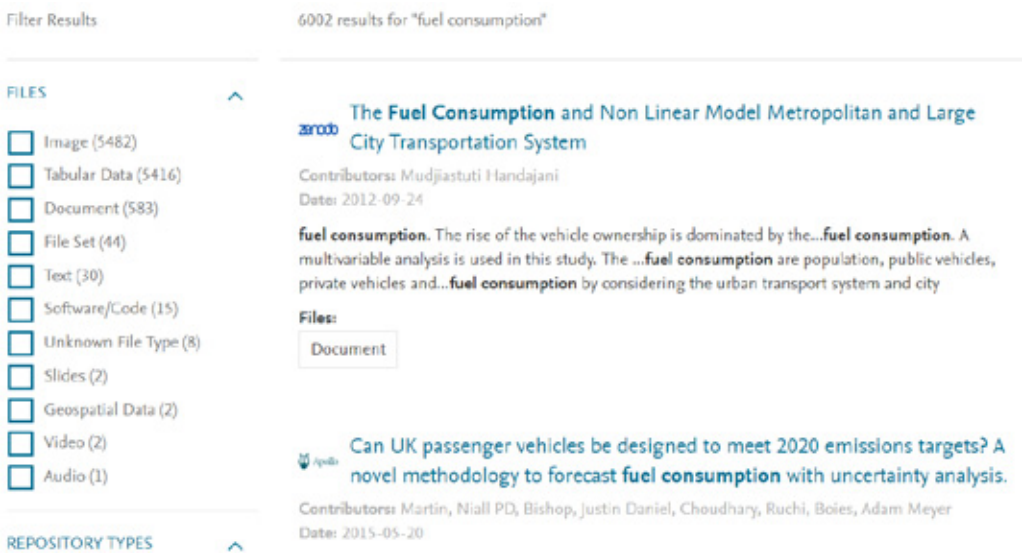


Źródło: https://www.mendeley.com/research-papers/?query=data-driven+ business+model

Oprócz samych publikacji Mendeley pozwala również na wyszukiwanie zbiorów danych. Funkcjonalność ta jest dostępna w osobnej zakładce 'Datasets'. Zgodnie z informacją ze strony można przeszukiwać spośród 10.3 miliona zbiorów danych. Co istotne, Mendeley pełni rolę agregatora, co oznacza, że możliwe jest wyszukiwanie nie tylko danych składowanych w portalu, ale również w zewnętrznych repozytoriach: 4TU, DRYAD, Dataverse, ICPSR, Zenodo, Apollo. Rys. **6.6** przedstawia wyniki wyszukiwania dla frazy 'fuel consumption'. Ujęcie frazy w cudzysłów oznacza wyszukanie takiej dokładnie sekwencji – otrzymujemy 6002 wyniki[21]. W dalszej kolejności można filtrować wg typu zasobu, np. wybrać tylko 'Tabular Data'. Rysunki można podejrzeć po kliknięciu wybranego zbioru (rys. **6.7**). Lista plików jest dostępna w podglądzie szczegółów zbioru (rys. **6.8**).

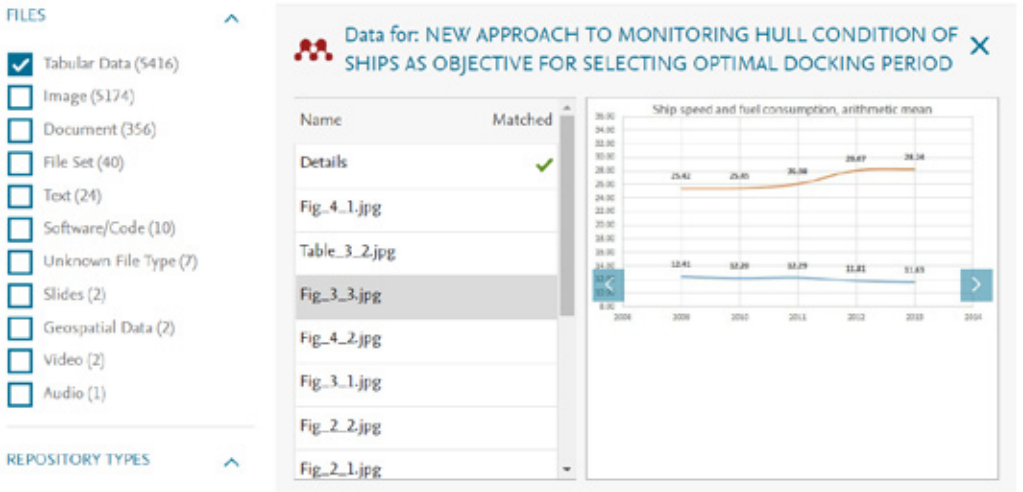
21 Badanie przeprowadzono 6.02.2019 r.

RYSUNEK 6.6. **WYNIKI WYSZUKIWANIA ZBIORÓW DANYCH W PORTALU MENDELEY**



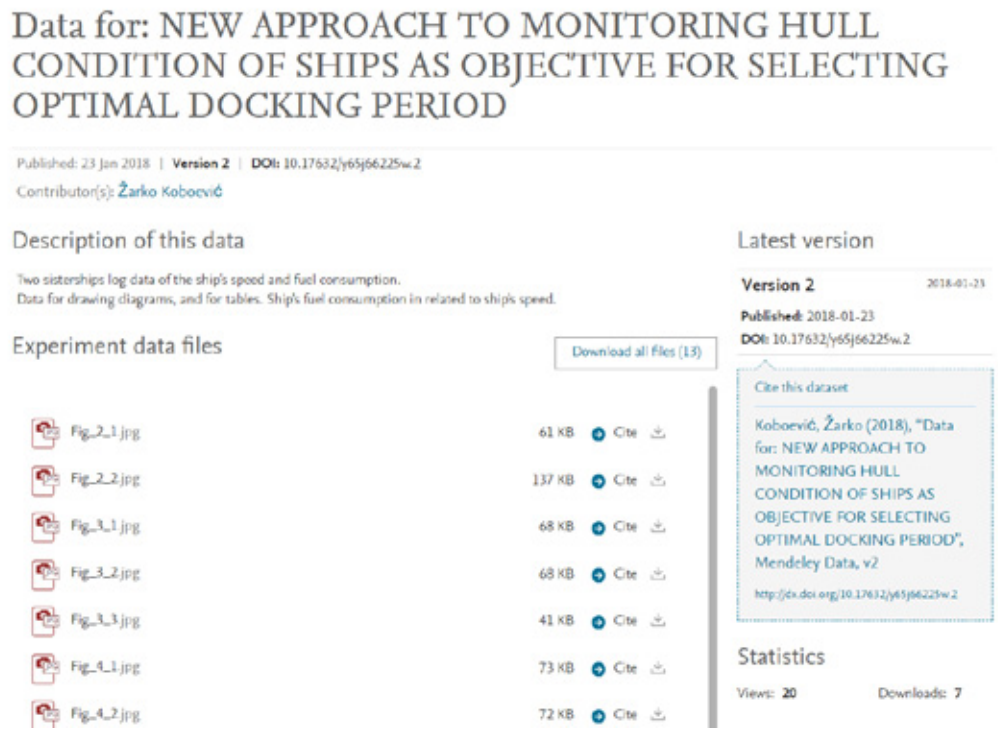
Źródło: https://data.mendeley.com/datasets?query=fuel+consumption

RYSUNEK 6.7. **PODGLĄD RYSUNKU PRZY WYNIKACH WYSZUKIWANIA W MENDELEY**



Źródło: https://data.mendeley.com/datasets?query=%22fuel+consumption%22&page=0&type=TABULAR_DATA

RYSUNEK 6.8. **PODGLĄD LISTY PLIKÓW W SZCZEGÓŁACH ZBIORU MENDELEY**



Źródło: https://data.mendeley.com/datasets/y65j66225w/2

6.4.2. ResearchGate

ResearchGate[22] jest w istocie siecią społecznościową dla naukowców. Możliwe jest anonimowe przeglądanie zawartości portalu, jednak tworzenie własnego profilu wymaga już rejestracji. Według danych na styczeń 2019 zarejestrowanych było ponad 15 milionów osób. Według Times Higher Education jest to portal dla naukowców z największą liczbą aktywnych użytkowników [**115**].

ResearchGate skupia się przede wszystkim na procesie promocji wyników badań. Można dodać metadane o publikacji we własnym profilu, z możliwością wgrania również pełnego tekstu. Poszczególne prace można organizować według projektów poprzez utworzenie dla nich specjalnych podstron. Można śledzić określonego naukowca i być śledzonym. RG dostarcza

22 https://www.researchgate.net/

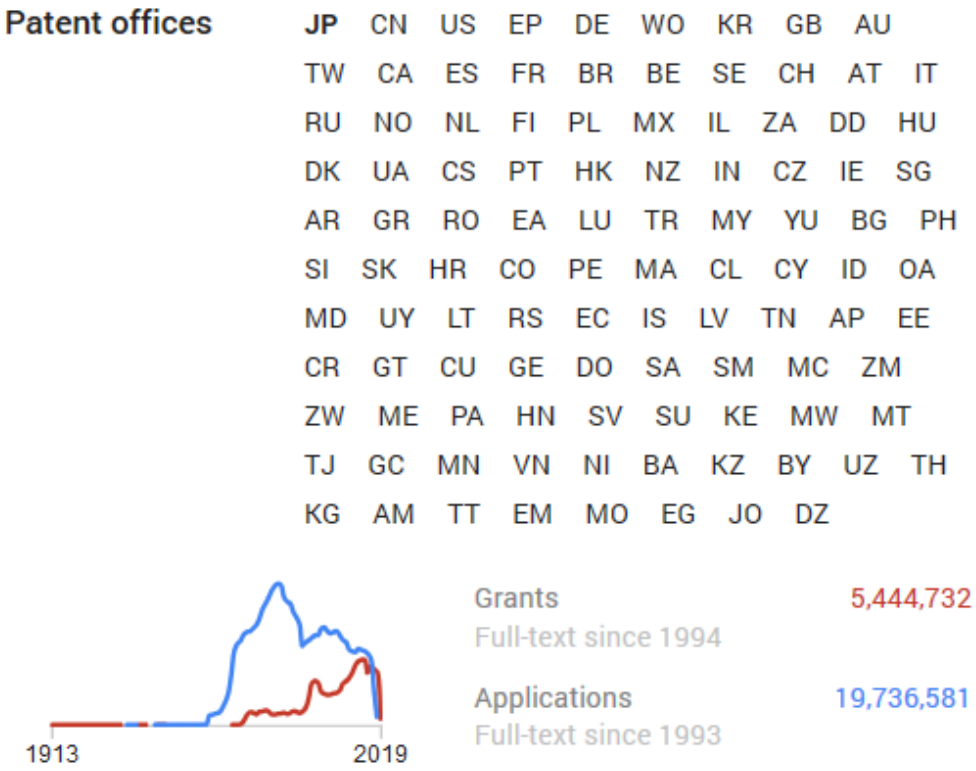
również statystyki dotyczące liczby odsłon danego artykułu oraz liczby cytowań. Poprzez uczestnictwo w dyskusji można zyskać dodatkowe punkty do scoringu naukowca.

Ze względu na brak funkcjonalności dołączania dodatkowych plików portal ten nie jest przydatny do poszukiwania danych badawczych. Nawet jeśli takie dane są dołączane jako dodatkowe materiały, to trudno je znaleźć ze względu na brak odpowiednich narzędzi.

6.4.3. Google patents

Google Patents[23] jest to doskonały portal do badania stanu techniki. Google indeksuje informacje o patentach oraz zgłoszeniach patentowych ze 106 krajowych biur patentowych (rys. **6.9**).

RYSUNEK 6.9. **BIURA PATENTOWE NA ŚWIECIE ORAZ STATYSTYKI DLA JAPONII W GOOGLE PATENTS**

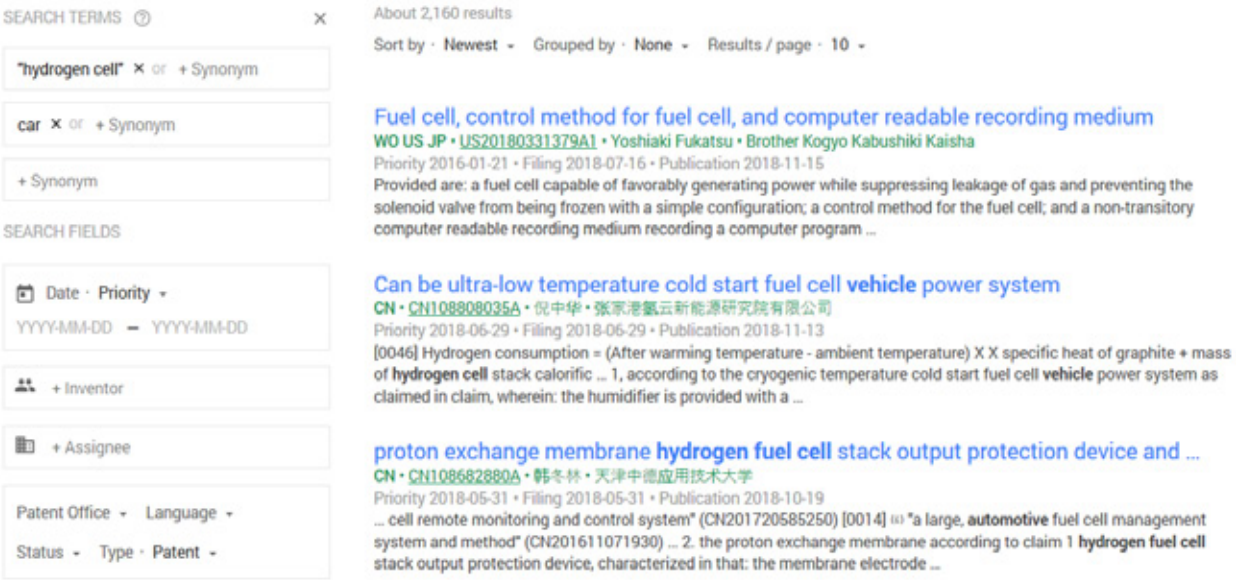


Źródło: https://patents.google.com/coverage/JP

23 https://patents.google.com/

Wyszukiwanie odbywa się w sposób intuicyjny poprzez wpisanie słów kluczowych. Rys. **6.10** przedstawia wyniki wyszukiwaniafrazy «„hydrogen cell" car». Dodatkowo jako typ dokumentów wybrane zostały patenty, a wyniki posortowane od najnowszych.

RYSUNEK 6.10. **WYNIKI WYSZUKIWANIE FRAZY „HYDROGEN CELL" W GOOGLE PATENTS**

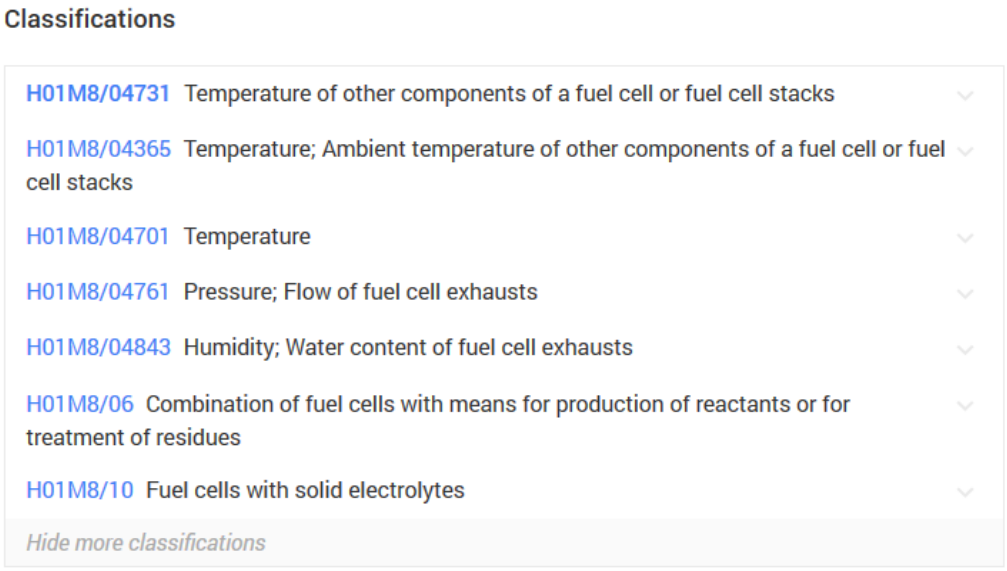


Źródło: zrzut ekranu.

Po przejściu na stronę poszczególnych patentów lub zgłoszeń patentowych jest możliwość zapoznania się z pełną treścią zgłoszenia patentowego (zastrzeżenia, ilustracje) oraz z metadanymi (rys. **6.12**). Z punktu widzenia badania stanu techniki istotne są również klasyfikacje patentowe, które również można znaleźć wśród metadanych (rys. **6.11**). Po klasyfikacjach można nawigować poprzez kliknięcie w odpowiedni link.

Oferowane są również dodatkowe funkcjonalności umożliwiające nawigację w bazie patentów. Można przejść bezpośrednio do podobnych dokumentów ('similar documents') – ich lista pojawia się na końcu opisu patentu. Dla niektórych patentów można znaleźć patenty cytowane ('Patent Citations') oraz cytujące ('Cited By'). Kliknięcie w nazwę zgłaszającego powoduje przejście do listy zgłoszeń i patentów.

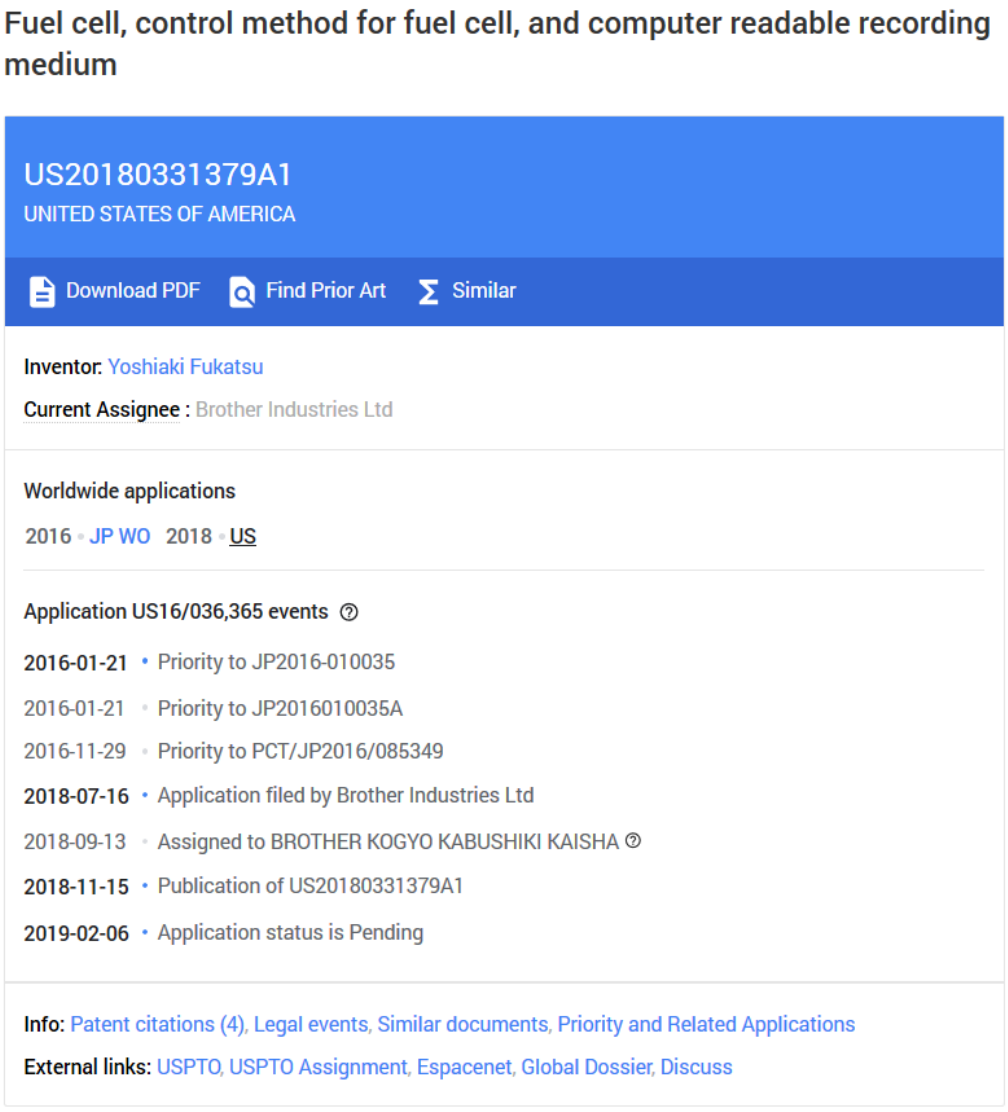
RYSUNEK 6.11. **PODGLĄD KLASYFIKACJI PATENTOWEJ W GOOGLE PATENTS**



Źródło: zrzut ekranu.

Możliwe jest również wyszukiwanie pośrednie. Pod numerem patentu na niebieskim pasku (rys. **6.12** u góry) znajdują się dwa przyciski. 'Find Prior Art' – wybiera automatycznie słowa kluczowe i poszukuje patentów najlepiej pasujących do takiego zapytania oraz dodatkowo jako kryterium ograniczające wskazana jest data pierwszeństwa. 'Similar' – powoduje wyszukanie dokumentów według podobieństwa pełnych tekstów patentów.

RYSUNEK 6.12. **SZCZEGÓŁY PATENTU W GOOGLE PATENTS**



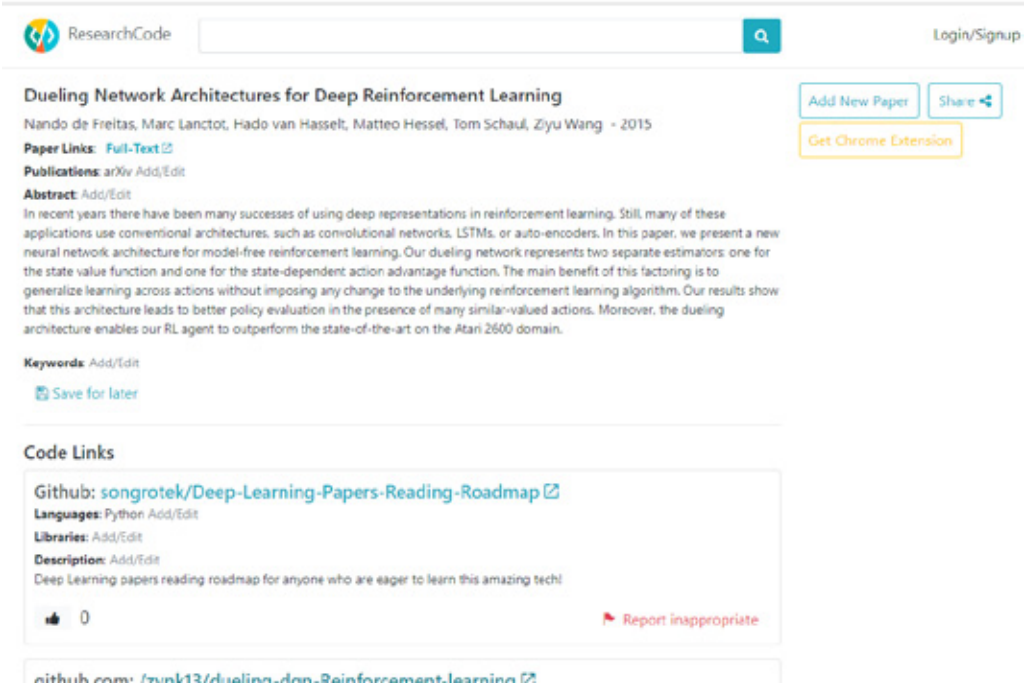
Źródło: zrzut ekranu.

6.4.4. ResearchCode – kod źródłowy

Ciekawą inicjatywą jest nowo powstała komercyjna usługa ResearchCode, pozwalająca na znalezienie kodu źródłowego przypisanego do danej publikacji naukowej. Usługa powstała dość niedawno, więc nie można jasno określić jej przydatności – ale pozostaje ciekawym źródłem dla osób zainteresowanych odtworzeniem przebiegu badań naukowych z wykorzystaniem zasobów danych (w szczególności w zakresie metod uczenia maszynowego).

Platforma ResearchCode pozwala połączyć publikacje z kodem źródłowym autora/ów umieszczonym w serwisie Github[24]. Rozwiązanie, wykorzystując web scraping i dane na temat repozytorium charakteryzuje (wykorzystując obecne w serwisie github dane) wykorzystany język programowania czy biblioteki. Kod powiązany z publikacjami opiera się na linkach i referencjach obecnych w samym repozytorium – przykładowe przetworzone przez serwis elementy znajdują się na rysunku **6.14**.

RYSUNEK 6.13. **PRZYKŁAD PUBLIKACJI Z WYLISTOWANYMI KODAMI ŹRÓDŁOWYMI W USLUDZE RESEARCHCODE**



Źródło: https://researchcode.com/code/1879375880/dueling-networkarchitectures-for-deep-reinforcement-learning/

24 www.github.com

RYSUNEK 6.14. **PRZYKŁAD PUBLIKACJI Z WYLISTOWANYMI KODAMI ŹRÓDŁOWYMI W USLUDZE RESEARCHCODE**

Źródło: www.github.com

6.5. Źródła danych

To, co z punktu widzenia przedsiębiorców jest najbardziej interesujące, to dane. Oczywiście często zrozumienie danych nie jest możliwe bez kontekstu, który może być dostarczony przez publikacje naukowe je wykorzystujące. W niniejszej sekcji przedstawiamy dostępne zasoby powstałe przede wszystkim w duchu koncepcji otwartej nauki. Są przedstawione tutaj najbardziej wartościowe repozytoria, tworzone zarówno przez społeczności, jak i podmioty komercyjne.

6.5.1. Zbiory danych do uczenia maszynowego

Zbiory do uczenia maszynowego mają pewne dodatkowe wymaganie, które sprawia, że są szczególnie poszukiwane przez przedsiębiorców, a jednocześnie trudne do znalezienia. Chodzi o to, że muszą zawierać „poprawne odpowiedzi", tj. jeśli mają być budowane modele do klasyfikacji, to poprawna klasyfikacja musi być znana; jeśli mają być tworzone modele predykcyjne, to oczekiwana prognoza modelu również musi być znana. Odpowiedzi mogą pochodzić z rzeczywistych danych. Często jednak jest tak, że muszą być ręcznie przygotowanie i ocenione przez człowieka, aby miały wartość z punktu widzenia uczenia modeli. s Jednym z ciekawszych źródeł, a jednocześnie wzorcowym opracowaniem, jest specjalna strona na Wikipedii poświęcona zbiorom danych związanymi z badaniami wykorzystującymi uczenie

maszynowe[25]. W istocie jest to lista zbiorów, ale jeden z aspektów przesądza o jej klasyfikacji w kategorii 'open science' – każdy zbiór opatrzony jest listą publikacji, w których był wykorzystany (przykład na rys. **6.15**). Zbiory zostały pogrupowane według zadań uczenia maszynowego i obejmują: dane obrazowe (np. rozpoznawanie twarzy, zdjęcia lotnicze), dane tekstowe (np. recenzje, tweety), dane dźwiękowe (np. mowa, myzyka), sygnalizacja (np. sygnały elektryczne), dane fizyczne (np. fizyka wysokich energii, astronomia), biologiczne (np. rośliny, zwierzęta, leki). Są również pojedyncze zbiory dedykowane do wykrywania anomalii oraz odpowiadania na pytania. Uzupełniają je dane wielowymiarowe, w szczególności obejmujące: finanse, pogoda, statystyka ludności, transport, internet, gry.

Istotnym zidentyfikowanym przez nas problemem jest jakość kolekcji zbiorów danych: często linki są nieaktualne, a z samego opisu często jest trudno wnioskować o jakości samego zbioru. Tym ważniejsze są inicjatywy, które nastawione są na zapewnienie aktualności list (tworzenie tzw. *curated repositories*), podobnie jak kuratorzy wystaw w muzeach. Zbiory danych mogą pochodzić z bardzo różnych źródeł i być reprezentowane w różnych formatach. Są społeczności, które podjęły wysiłek uporządkowania i standaryzacji dostępu do zbiorów, tak aby łatwiej było je wykorzystać w badaniach związanych z uczeniem maszynowym. Najbardziej prominentne przykłady to OpenML oraz PLMB.

6.5.2. OpenML

OpenML [**179**] jest platformą webową, która zapewnia dostęp do tysięcy zbiorów danych w zunifikowany sposób, pozwalając na ewaluację i benchmarking algorytmów. Wiele odkryć dokonano m.in. dzięki wdrożeniu narzędzi do uporządkowanego zarządzania danymi, zapewniających powtarzalność i porównywalność analiz. Narzędzia takie dostępne on-line są ciekawym uzupełnieniem teoretycznych artykułów publikowanych w czasopismach. Zgodnie z określeniem twórców, OpenML jest miejscem dla badaczy zajmujących się uczeniem maszynowym, gdzie mogą organizować i współdzielić dane wykorzystane w badaniach, ułatwiając współpracę z innymi badaczami.

Po wejściu na stronę główną[26] można przeczytać informację o podstawowych statystykach. Łącznie na platformie przechowywane jest 20.249 zbiorów (w tym 2546 zweryfikowanych i aktywnych), realizowane jest 68.234 zadań naukowych oraz zdefiniowano 6254 przepływy analityczne[27]. Łącznie zadania eksploracji uruchamiano ponad 9,7 mln razy.

25 https://en.wikipedia.org/wiki/List_of_datasets_for_machine_ learning_research
26 https://www.openml.org/
27 Stan na 5.02.2019 r.

RYSUNEK 6.15. **PRZYKŁADOWA TABELKA Z LISTĄ ZBIORÓW WYKORZYSTYWANYCH W UCZENIU MASZYNOWYM**

Facial recognition [edit]

In computer vision, face images have been used extensively to develop facial recognition systems, face detection, and many other projects that use images of faces.

Dataset name	Brief description	Preprocessing	Instances	Format	Default task	Created (updated)	Reference	Creator
FERET (facial recognition technology)	11338 images of 1199 individuals in different positions and at different times.	None.	11,338	Images	Classification, face recognition	2003	[6][7]	United States Department of Defense
CMU Pose, Illumination, and Expression (PIE)	41,368 color images of 68 people in 13 different poses.	Images labeled with expressions.	41,368	Images, text	Classification, face recognition	2000	[8][9]	R. Gross et al.
Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)	7,356 video and audio recordings of 24 professional actors. 8 emotions each at two intensities.	Files labelled with expression. Perceptual validation ratings provided by 319 raters.	7,356	Video, sound files	Classification, face recognition, voice recognition	2018	[10][11]	S.R. Livingstone and F.A. Russo

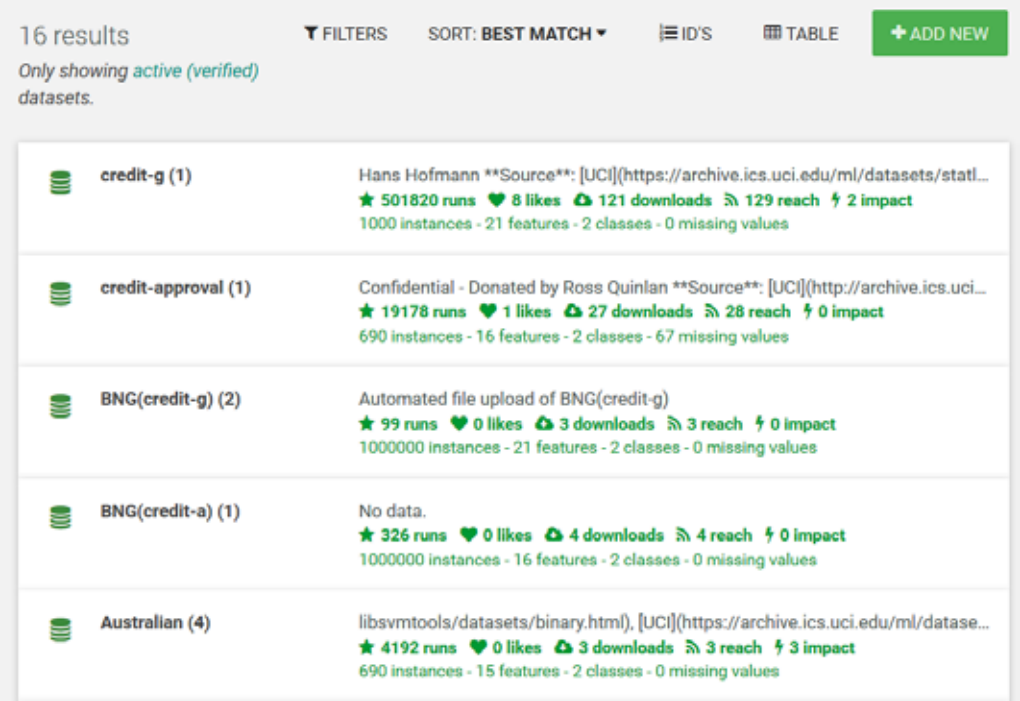
Źródło: https://en.wikipedia.org/wiki/List_of_datasets_for_machine_ learning_research

Dane można wyszukiwać po nazwie oraz sortować wyniki według wielu kryteriów. Rysunek **6.16** przedstawia wyniki wyszukiwania według frazy 'credit'. Każdy zbiór opatrzony jest statystykami, które mogą wskazywać na przydatność tego zbioru z perspektywy innych użytkowników, np. liczba uruchomień, polubień, ściągnięć zbioru, zasięg. Są również informacje typowo techniczne dotyczące np. wielkości zbioru: liczba instancji, cech oraz brakujących danych, a w przypadku zbioru do klasyfikacji – również liczba wyróżnionych klas.

Po wejściu na stronę danego zbioru można podejrzeć jego szczegóły. Pokazywane są przede wszystkim metadane pozwalające na ocenę, czy zbiór jest przydatny (rys. **6.17**) oraz rozkład poszczególnych cech czy zmiennych celu (rys. **6.18**).

Istotną zaletą OpenML jest możliwość integracji z wieloma środowiskami i narzędziami analitycznymi. Oferowane są API REST-owe, Python, R, Java, C#. Ściąganie danych może odbywać się bez żadnych formalności. Ładowanie danych na portal wymaga jednak założenia konta użytkownika i uwierzytelnienia za każdym razem. OpenML jest również silnie zintegrowany z popularnym pakietem Pythona do data science – scikit-learn. Przykładowy kod włączenia zewnętrznego zbioru danych do własnej analizy został przedstawiony w listingu **6.1**. Poniższy kod nie tylko ściąga pożądane dane, ale również umożliwia zdefiniowanie przepływów danych w rozumieniu OpenML i umożliwia załadować tam wyniki uruchomienia.

RYSUNEK 6.16. **WYNIKI WYSZUKIWANIA FRAZY 'CREDIT' W PORTALU OPENML**

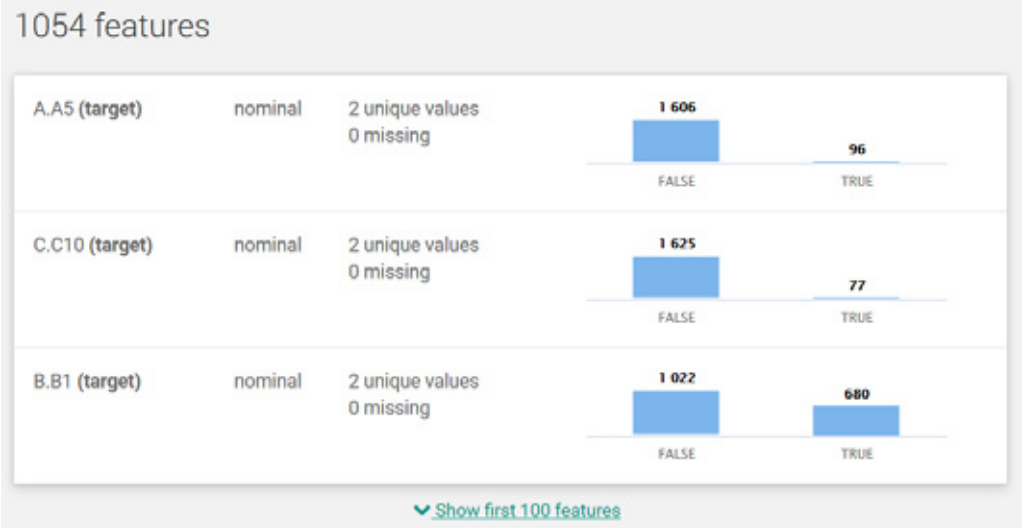


Źródło: https://www.openml.org/search?q=enron&type=data

RYSUNEK 6.17. **OPENML – PODGLĄD METADANYCH ZBIORU**

Źródło: https://www.openml.org/d/40590

RYSUNEK 6.18. **OPENML – PODGLĄD ROZKŁADU WARTOŚCI ZMIENNEJ CELU**



Źródło: https://www.openml.org/d/40590

KOD ŹRÓDŁOWY 6.1. **KOD INTEGRACJI OPENML W PYTHONIE**

```
from sklearn import ensemble
from openml import tasks, flows, Runs

task = tasks.get_task(1234)
clf = ensemble.RandomForestClassifier()
flow = flows.sklearn_to_flow(clf)
run = runs.run_flow_on_task(task, flow)
result = run.publish()
```

6.5.3. PMLB

Drugim wartym uwagi konserwowanym repozytorium danych jest PMLB (Penn Machine Learning Benchmark) [**133**]. Jak nazwa wskazuje, jest to głównie repozytorium zbiorów danych służących ocenie algorytmów (benchmark). Zawiera zbiory do klasyfikacji oraz regresji w standaryzowanym formacie, dostępne przez API Pythona.

Nie ma odrębnego interfejsu webowego, zbiory danych są umieszczone bezpośrednio w repozytorium github[28]. Stąd też wyszukanie odpowiedniego zbioru jest znacznie utrudnione. Wprawdzie dla zbiorów dostępny jest opis to ogranicza się on do podania liczby instancji, cech,

28 https://github.com/EpistasisLab/penn-ml-benchmarks/tree/master/datasets

klas itp., określone są również typy zmiennych. Dokumentacja dostępna jest tylko dla części zbiorów (gł. tych od klasyfikacji), nie jest dostępna dla zbiorów do regresji. Nie jest wskazane, z jakiego źródła pochodzi dany zbiór. Same nazwy zbiorów są dość enigmatyczne, gdzie obok siebie mamy: car, cars, cars1.

Na repozytorium Penn Machine Learning Benchmark (PMLB) składają się zbiory pobrane z innych kolekcji – znanych benchmarków dla uczenia maszynowego, w tym: UCI ML repository[29], Kaggle[30], KEEL[31]. Według twórców PMBL zawiera większość benchmarków stosowanych w obszarze studiów nad ML. Obecnie[32] dostępne jest 166 zbiorów do klasyfikacji oraz 120 zbiorów na potrzeby regresji.

Aby repozytorium zbiorów było łatwe w użyciu, wszystkie umieszczone w nim dane zostały wstępnie przetworzone. Dane są zapisywane jako pliki tekstowe rozdzielane znakiem tabulacji (TSV), zmienne umieszczone są w kolumnach, instancje w wierszach, pierwszy wiersz jest zawsze nagłówkiem kolumn. Zmienna zależna została oznaczona jako 'class' dla klasyfikacji i 'target' dla regresji. Dodatkowo wszystkie zmienne kategoryczne oraz etykietki, które były nienumeryczne, zostały zamienione na numeryczne odpowiedniki (np. „Low", „Medium", „High" zostały zastąpione przez 0, 1, 2). Tutaj też ujawnia się pewien mankament tych zbiorów – nie zostało udokumentowane, w jaki sposób te zmienne zostały podstawione, tzn. nie ma możliwości odtworzenia danych pierwotnych. Co więcej, przy wczytywaniu zbioru w Pythonie nie jest oczywiste, które zmienne są kategorialnymi, a które liczbowymi. Taka informacja znajduje się tylko na wiki tego projektu i w oczywisty sposób nie jest przetwarzalna maszynowo. Nawet jeśli przyjąć, że chodzi tylko o benchmark algorytmów do uczenia maszynowego, to takie algorytmy powinny mieć informacje o tym czy zmienna była wcześniej kategorią czy liczbą.

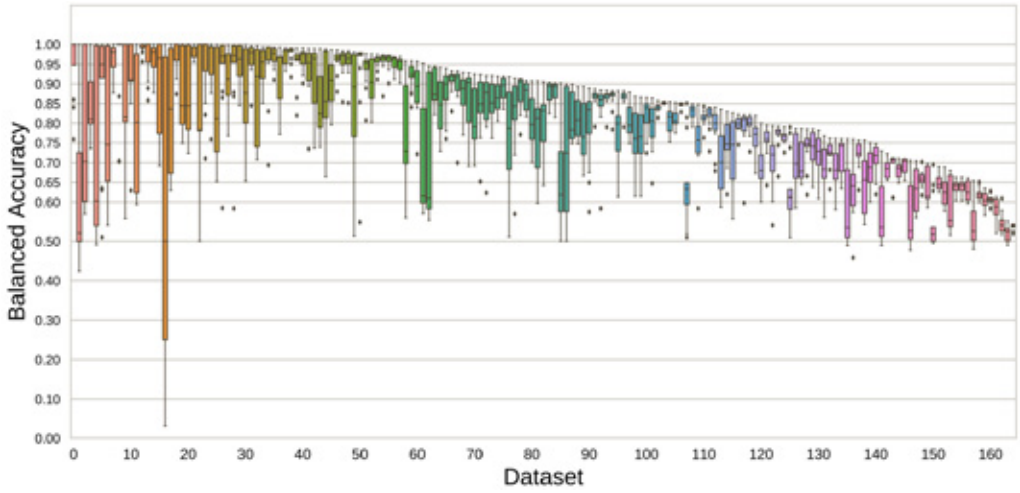
Głównym celem było ułatwienie korzystania ze zbiorów w Pythonie, dlatego też został dostarczony odpowiedni pakiet. Przykład włączenia zbiorów do kodu w Pythonie został przedstawiony w listingu **6.2**. Więcej szczegółów oraz wskazówek programistycznych zostało umieszczonych na stronie wiki odpowiedniego projektu na GitHubie[33]. Możliwe jest również wylistowanie wszystkich dostępnych zbiorów. Dzięki temu trywialne staje się przeprowadzenie benchmarku algorytmu i przygotowanie profilu porównującego jakość dopasowania modelu na wszystkich zbiorach (zob. rys. **6.19**).

29 http://archive.ics.uci.edu/ml/
30 http://www.kaggle.com/
31 http://www.keel.es/
32 Stan na 5.02.2019 r., ostatnia aktualizacja 13.02.2018 r.
33 https://github.com/EpistasisLab/penn-ml-benchmarks

KOD ŹRÓDŁOWY 6.2. **KOD INTEGRACJI PMLB W PYTHONIE**

```
from pmlb import fetch_data

# Returns a pandas DataFrame
adult_data = fetch_data('adult')
print(adult_data.describe())
# Returns NumPy arrays
adult_X, adult_y = fetch_data('adult', return_X_y=True)
print(adult_X)
print(adult_y)
```

RYSUNEK 6.19. **DOKŁADNOŚĆ DOSTROJONYCH MODELI ML NA KAŻDYM ZBIORZE Z REPOZYTORIUM PMLB, SORTOWANIE WG MALEJĄCEJ DOKŁADNOŚCI**



Źródło: [133]

6.5.4. Dataverse

Dataverse jest oprogramowaniem rozwijanym przez Institute for Quantitative Social Science (IQSS), afiliowany przy Unwersytecie Harvarda, we współpracy z wieloma naukowcami ze świata[34]. Zostało ono rozwinięte na podstawie wcześniejszych doświadczeń Centrum Danych Harvard-MIT oraz Biblioteki Uniwersytetu Harvarda w tym obszarze, zdobytych podczas realizacji projektu Virtual Data Center (VDC) w latach 1999-2006.

34 https://dataverse.org/about

Głównym celem oprogramowania Dataverse jest zarządzanie repozytoriami danych. Funkcjonuje ono jako aplikacja webowa, która umożliwia zachowanie, współdzielenie, cytowanie, eksplorację i analizę danych badawczych. Jako główni odbiorcy rozwiązania są wskazywani: naukowcy – zachęta do zwiększenia widoczności własnych badań i cytowalności, czasopisma – powiązanie danych z publikowanymi artykułami je wykorzystującymi, instytucje – usprawnienie zarządzania danymi, w tym również ich utrwalanie i integracja z innymi zbiorami, deweloperzy – rozwój platformy. Nie są tutaj wymienieni faktycznie korzystający z platformy, niemniej oczywiste jest, że poszukujący danych, czy to osoby indywidualne, czy przedsiębiorstwa, są w istocie głównymi odbiorcami tego rozwiązania.

Ideą Dataverse jest usprawnienie publicznej dystrybucji utrwalonych, autoryzowanych i weryfikowalnych danych, wsparte przez zaawansowaną, ale łatwą w użyciu technologię, która umożliwia pracę również z poufnymi lub prywatnymi danymi. Pojedyncza instalacja oprogramowania określana jest jako repozytorium. W jej ramach można tworzyć wiele wirtualnych archiwów określanych jako 'dataverses' (na wzór 'universe'). Każdy dataverse zawiera zbiory danych ('datasets'), a te z kolei opisowe metadane i pliki, w tym pliki z danymi, dokumentację, przykładowy kod źródłowy [**88**]. Każdy opublikowany zbiór otrzymuje swój numer DOI (Digital Object Identifier), dzięki czemu możliwe jest jego cytowanie.

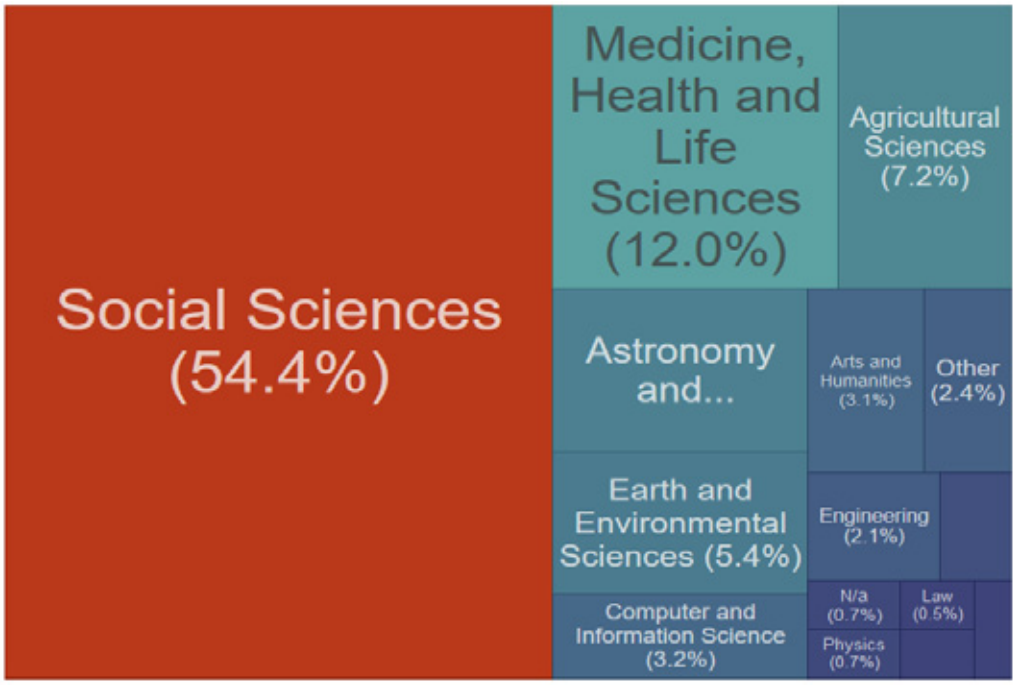
RYSUNEK 6.20. **INSTALACJE DATAVERSE PRZEDSTAWIONE NA MAPIE ŚWIATA**

Źródło: https://dataverse.org/

Pierwotnie rozwiązanie to skupiało się na danych w naukach społecznych. Obecnie zakres danych jest znacznie szerszy. Dataverse stosowane jest przez różne instytucje zajmujące się bardzo różnymi obszarami nauki, od ogólnych (np. Ministerstwo Innowacji w Brazylii) do bardziej szczegółowych, jak zdrowie publiczne czy rolnictwo (np. CIRAD – Agriculture Research for Development). Obecnie działa 36 instalacji Dataverse (rys. **6.20**).

Wśród instalacji pierwotną i obecnie najbardziej znaną jest Harvard Dataverse[35]. Znajduje się tam 3115 dataversów, 81.579 kolekcji danych (datasets) oraz 492.798 plików (files). Każdy dataverse ma przypisaną swoją kategorię oraz klasyfikację przedmiotową, co ułatwia wyszukiwanie określonych danych. Najwięcej jest dataversów poświęconych projektom badawczym (1053) oraz tworzonych przez poszczególnych badaczy (928). Najwięcej jest kolekcji danych poświęconych naukom społecznym – 37.539, co w kontekście wszystkich repozytoriów stanowi ponad 50% zbiorów (zob. rys. **6.21**).

RYSUNEK 6.21. **KLASYFIKACJA PRZEDMIOTOWA ZBIORÓW ZEBRANYCH PRZEZ DATAVERSE**



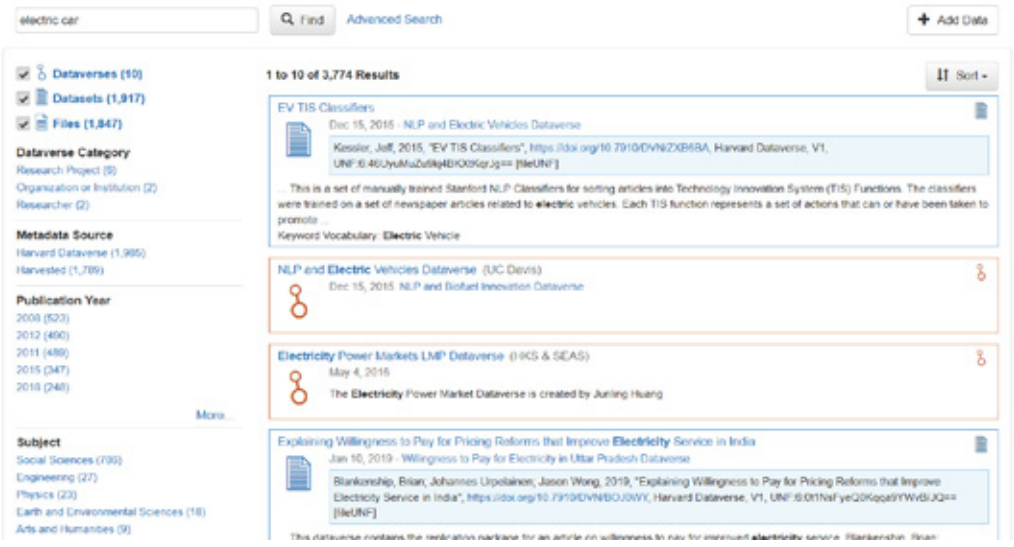
Źródło: https://dataverse.org/metrics

35 https://dataverse.harvard.edu

Załóżmy, że interesujemy się samochodami elektrycznymi i chcemy sprawdzić, czy są jakieś dane dotyczące tego zagadnienia. W polu wyszukiwania wpisujemy słowa kluczowe 'electric car'. Otrzymujemy 3774 wyniki, w poszczególnych typach zasobów (oznaczonych odpowiednimi ikonami i kolorami): 10 dataversów, 1917 kolekcji oraz 1847 plików (zob. rys. **6.22**). Dla zwiększenia precyzji można wpisać frazę w cudzysłowie, aby wyrazy te znalazły się obok siebie. Dużo wyników zostało zaklasyfikowanych w obszarze nauk społecznych, dlatego możemy użyć filtra dyscypliny ('Subject') i skupić się wyłącznie na danych z zakresu inżynierii.

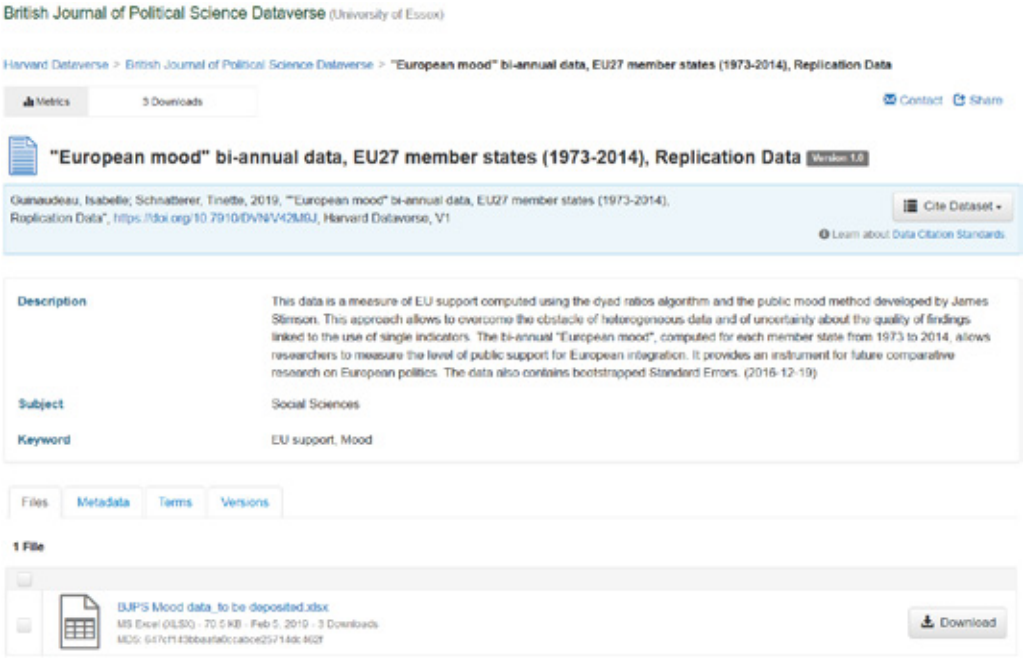
Po wejściu na stronę wybranej kolekcji otrzymujemy o niej informację (rys. **6.23**). Przedstawiony jest krótki opis, klasyfikacja oraz słowa kluczowe. W dolnej części można przejrzeć dołączone do kolekcji pliki, metadane oraz, co istotne przy ponownym użyciu danych, informację o licencji. Poszczególne pliki przedstawione na osobnych stronach (rys. **6.24**) i poddawane są niezależnej kontroli wersji.

RYSUNEK 6.22. **WYNIKI WYSZUKIWANIA FRAZY 'ELECTRIC CAR' W REPOZYTORIUM HARVARD DATAVERSE**



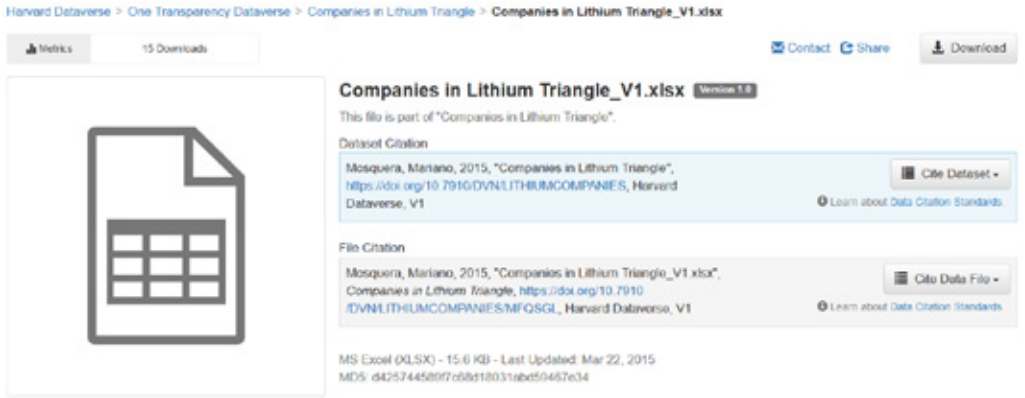
Źródło: https://dataverse.harvard.edu/dataverse/harvard?q=electric+car

RYSUNEK 6.23. **PRZYKŁADOWA KOLEKCJA DANYCH PUBLIKOWANA NA DATAVERSE WRAZ Z OPISEM**



Źródło: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/V42M9J

RYSUNEK 6.24. **PRZYKŁADOWY PLIK PUBLIKOWANY NA DATAVERSE**



Źródło: https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/LITHIUMCOMPANIES/MFQSGL&version=1.0

6.5.5. Zenodo

Zenodo[36] jest otwartym repozytorium ogólnego przeznaczania rozwiniętym w ramach europejskiego programu OpenAIRE i obsługiwane przez CERN. Umożliwia badaczom składowanie własnych zbiorów danych, oprogramowania, raportów i innych artefaktów związanych z prowadzonymi badaniami. Każde zgłoszenie jest opatrzone jednoznacznym identyfikatorem (DOI), co czyni je łatwo cytowalnym. Zenodo również odnosi się do zasad FAIR[37] – zbiory powinny być znajdowalne (F), dostępne (A), interoperacyne (I) oraz gotowe do wielokrotnego wykorzystania (R).

Zenodo zostało uruchomione w maju 2013 z inicjatywy Komisji Europejskiej, której zależało na tym, aby w jednym miejscu składować wszystkie wyniki finansowanych przez nią projektów. Na składowanie danych badaczy przeznaczony jest klaster dysków o rozmiarze 18 petabajtów. Wpierając swoje programy badawcze CERN rozwinął narzędzia do zarządzania big data i rozszerzył możliwości bibliotek cyfrowych w kierunku danych otwartych. Oprogramowanie dostępne jest na otwartej licencji, a jego kod źródłowy znajduje się w dedykowanym repozytorium[38].

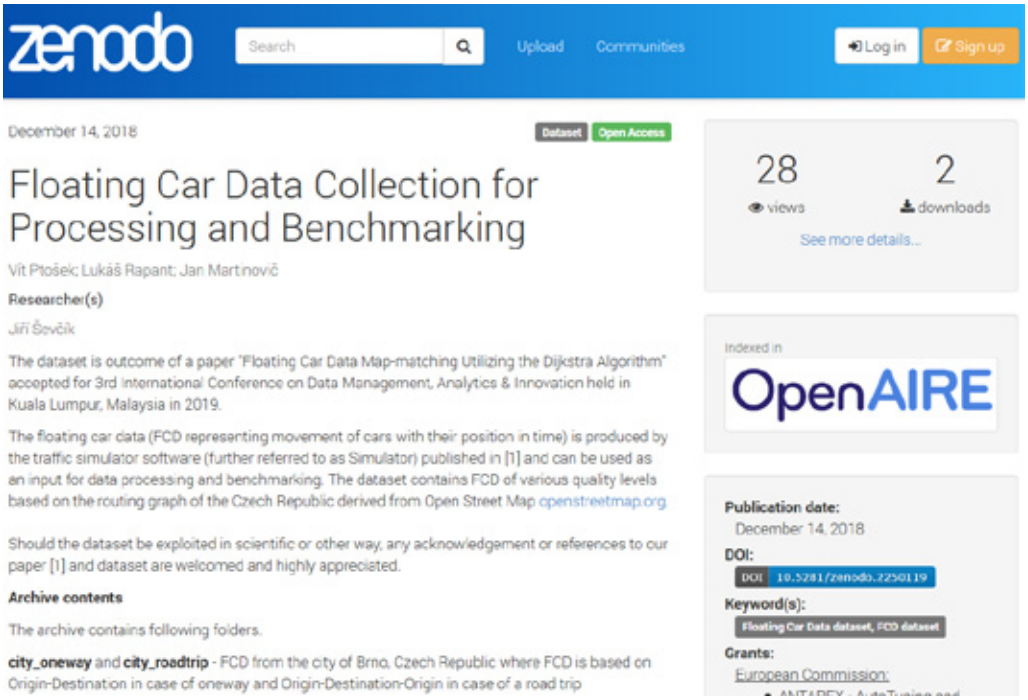
Zenodo kładzie nacisk na prostotę obsługi i narzędzi. Powinny one być używalne przez naukowców z dowolnej dziedziny. W repozytorium można składować dowolny artefakt, niezależnie od formatu pliku, sposobu dostępu, licencji etc. Najważniejsze typy artefaktów, po których można później filtrować wyniki zapytań, to publikacje, zbiory danych, obrazy, oprogramowanie (również bezpośredni import z githuba), prezentacje czy poster.

Jeśli chodzi o wyszukiwanie, to dostępny jest tylko jeden sposób wyszukiwanie – proste wpisanie słów kluczowych w pasku wyszukiwania. Sam portal nie jest specjalnie rozbudowany od strony graficznej. Przykładowa prezentacja zbioru przedstawiona jest na rys. **6.25**. Metadane opisujące zbiór danych są udokumentowane w formacie tekstowym, nie są przetwarzalne maszynowo.

Pomocniczym sposobem na wyszukiwanie danych jest korzystanie ze społeczności, których na Zenodo zostało utworzonych 2709. Społeczności często grupują zbiory danych tematycznie. Jednym z przykładów jest społeczność 'Wind energy' przedstawiona na rys. **6.26**. Na rysunku można również zauważyć integrację z ORCID, jeśli do zbiorów przypisani są autorzy. Oznacza to, ze tak opublikowane zbiory danych mogą być automatycznie dołączone do profili badawczych ORCID.

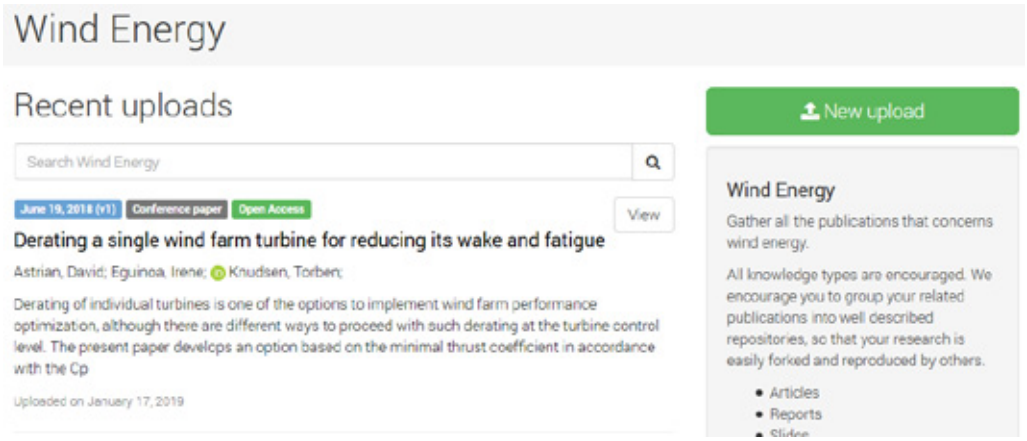
36 https://zenodo.org/
37 http://about.zenodo.org/principles/
38 https://github.com/zenodo/zenodo

RYSUNEK 6.25. **PREZENTACJA ZBIORU W REPOZYTORIUM ZENODO**



Źródło: https://zenodo.org/record/2250119

RYSUNEK 6.26. **ZBIORY DANYCH DOSTĘPNE W RAMACH SPOŁECZNOŚCI W ZENODO**



Źródło: https://zenodo.org/communities/wind_energy/

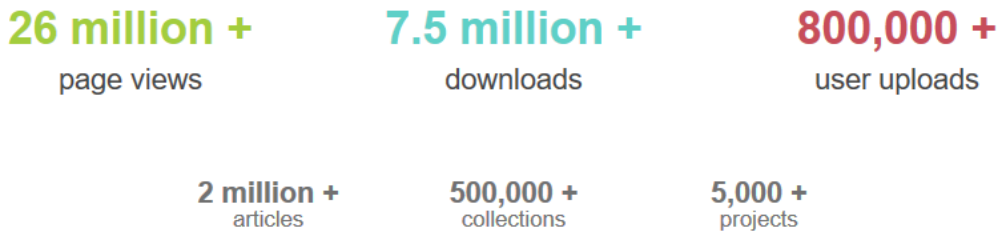
6.5.6. Figshare

Figshare[39] to kolejny portal umożliwiający umieszczanie, a tym samym również znajdowanie zbiorów danych wykorzystywanych w badaniach. Wspierany jest przez wiele wydawnictw na świecie, a wśród nich również największych graczy: Palgrave Macmillan, Wiley, PLOS, The Royal Society, Taylor & Francis, Springer Nature.

Obecnie jako typ danych można publikować: ilustracje, zbiory danych, dane multimedialne, artykuły (również preprinty), postery, kod źródłowy i grupy plików. Możliwe jest publikowanie zbiorów o rozmiarze do 5GB, a każdy użytkownik otrzymuje darmowe 20GB prywatnej przestrzeni; przestrzeń publiczna jest nielimitowana. Co jest istotne szczególnie dla osób poszukujących zbiorów do swobodnego wykorzystania, wszystkie pliki publikowane są na licencji Creative Commons (CC).

Przepływ danych w procesie publikacji został określony przez następujące kroki: Upload – wgranie pliku, Manage – zarządzanie nim, Share – współdzielenie, Publish – udostępnienie społeczności. Podobnie jak w przypadku wcześniej opisanych rozwiązań, publikowane zbiory otrzymują swój identyfikator DOI, dzięki temu możliwa jest ich jednoznaczna identyfikacja oraz możliwe jest cytowanie. Twórcy portalu jako motywację podają statystyki cytowań – treści z Figshare do 1.02.2019 r. cytowane były ponad 5000 razy. Statystyki przedstawione są na rys. **6.27**.

RYSUNEK 6.27. **SATYSTYKI FIGSHARE ZEBRANE OD 2012 ROKU**



Źródło: https://figshare.com/about

39 https://figshare.com/

Figshare jest powiązane z innymi portalami, m.in.ORCID[40], Symplectic Elements[41], może importować elementy z Github[42], a statystyki cytowań danych są zbierane przez Altmetric[43].

Na rysunku **6.28** przedstawiono kategorie danych, które można znaleźć na figshare. Ze względu na sposób prezentacji nie można znaleźć, ile plików jest dostępnych. Autorzy portalu również nie podają takich statystyk.

W celu znalezienia interesującego nas zbioru można skorzystać z wyszukiwania za pomocą słów kluczowych lub nawigować w strukturze kategorii zaprojektowanej przez twórców rozwiązania. W pierwszej kolejności należy określić główna kategorię (rys. **6.28**).

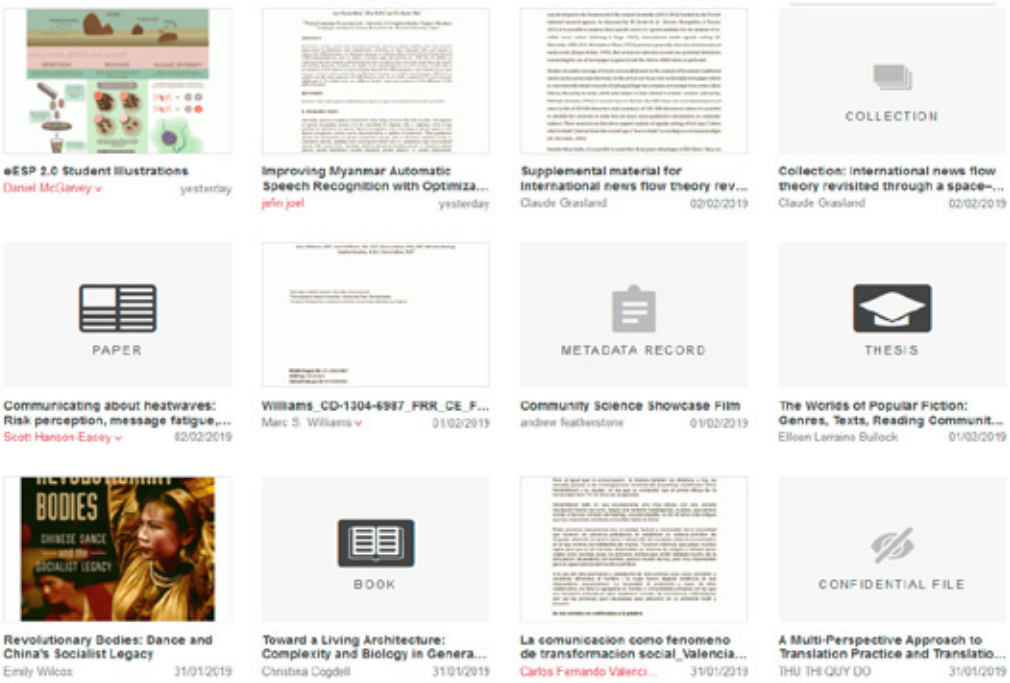
RYSUNEK 6.28. **GŁÓWNE KATEGORIE DANYCH W REPOZYTORIUM FIGSHARE**

Źródło: https://figshare.com/

Po wybraniu kategorii prezentowane są podkategorie (hierarchia jest dwustopniowa) oraz wyświetlane są zasoby w postaci kafelek (rys. **6.29**).

40 Baza naukowców, https://orcid.org/
41 Inne podobne rozwiązanie, https://symplectic.co.uk/products/elements-3/
42 https://github.com/
43 https://www.altmetric.com

RYSUNEK 6.29. **PREZENTACJA WYNIKÓW WYSZUKIWANIA W FIGSHARE**



Źródło: https://figshare.com/categories/Language_Communication_and_ Culture/1221

6.6. Analiza przypadków

W niniejszej sekcji opisujemy studia przypadków związane z wykorzystaniem danych dostępnych w ramach open science.

6.6.1. Badanie stanu techniki

Załóżmy, że szukamy informacji o nowych metodach napędzania pojazdów. Chcemy się dowiedzieć, jakie technologie są obecnie rozwijane oraz jaka jest ich popularność. Dużo się mówi o samochodach elektrycznych, ale być może postęp odbywa się również w napędach tradycyjnych.

W pierwszej kolejności należy ustalić klasyfikację patentową, w której będziemy poszukiwali rozwiązań. Na stronie Espacenet[44] znajdujemy Wspólną Klasyfikację Patentową, którą można przeszukać, wpisując odpowiednie słowa kluczowe. Na potrzeby naszego badania wybieramy 'vehicle engine'. Wyniki przedstawia rys. **6.30**. Zauważamy, że wiele wyników skupia się wokół klasy B60, która jest definiowana jako 'Vehicles in general'.

44 https://pl.espacenet.com/classification?locale=pl_PL

Analiza klasy B60 doprowadza nas do klasy B60K, w której sklasyfikowane są patenty dotyczące napędu (*propulsion*). Rys. **6.31** przedstawia rozwinięcie klasy B60K. W kolejnych podklasach znajdują się różne rodzaje napędów: elektryczne, parowe, wewnętrznego spalania, odrzutowe, również o wielu rodzajach napędów. W zakresie zainteresowania będą więc leżały: B60K 1/00, B60K 3/00, B60K 5/00, B60K 6/00, B60K 8/00.

RYSUNEK 6.30. **WYNIKI WYSZUKIWANIA 'VEHICLE ENGINE' NA ESPACENET**

Symbol	Symbol klasyfikacji i opis
★★★★★★★★★★ B60K 6/00	**Arrangement or mounting of plural diverse prime-movers for mutual or common propulsion, e.g. hybrid propulsion systems comprising electric motors and internal combustion engines** {; Control systems therefor, i.e. systems controlling two or more prime movers, or controlling one of these prime movers and any of the transmission, drive or drive units (arrangement or mounting in vehicles of electrical gearing, in which an electrical machine serves only as reduction gearing and not as the prime mover and in which no electrical storing means are used B60K 17/12; control and regulation of purely electrical prime movers B60L; prime-movers comprising electrical and internal combustion motors in a common engine block or housing per se F02B 65/00; electric motors or motor-generators used for starting the combustion engine F02N 11/04; electric motors for synchronising gearing F16H 3/12)}{Informative references: mechanical gearings with secondary electric drive F16H 3/72; arrangements for handling mechanical energy structurally associated with the dynamo-electric machine H02K 7/00; machines comprising structurally interrelated motor and generator parts H02K 51/00; dynamo-electric machines not otherwise provided for in H02K see H02K 99/00}
★★★★★★★★★★ B60W 10/00	**Conjoint control of vehicle sub-units of different type or different function** (for propulsion of purely electrically-propelled vehicles with power supplied within the vehicle B60L 50/00)
★★★★★★★★★★ B60W 20/00	**Control systems specially adapted for hybrid vehicles** {(hybrid vehicle design, B60K 6/00; electric vehicles B60L)}
★★★★★★★★★★ Y02T 10/00	**Road transport of goods or passengers**
★★★★★★★★★★ B60W 2710/00	**Output or target parameters relating to a particular sub-units**
★★★★★★★★★★ Y10S 903/00	**Hybrid electric vehicles, HEVS**
★★★★★★★★★★ B60W 2510/00	**Input parameters relating to a particular sub-units**
★★★★★★★★★★ B60Y 2200/00	**Type of vehicle** (not used; see subgroups)
★★★★★★★★★★ B60W 30/00	**Purposes of road vehicle drive control systems not related to the control of a particular sub-unit, e.g. of systems using conjoint control of vehicle sub-units** {, or advanced driver assistance systems for ensuring comfort, stability and safety or drive control systems for propelling or retarding the vehicle (anti-lock brake systems [ABS] B60T 8/00)}
★★★★★★★★★★ F02D 41/00	**Electrical control of supply of combustible mixture or its constituents** (F02D 43/00 takes precedence {; control of engine starters F02N 11/08, electrical control of engine ignition timing F02P 5/145)}

Źródło: https://pl.espacenet.com/classification?locale=pl_PL#!/q= vehicle%20engine

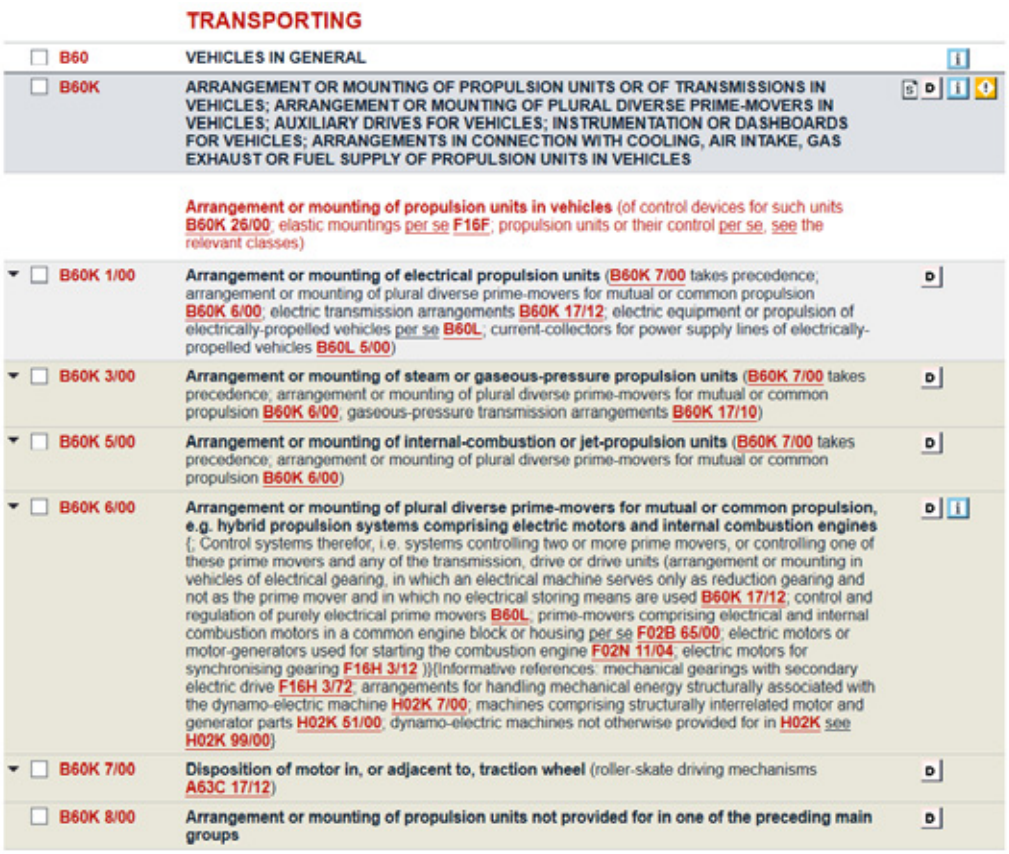
Korzystając z Google Patents można jako kryterium wyszukiwania wpisać nazwę klasy. Tabela **6.1** przedstawia liczbę wyników dla poszczególnych klas. Najwięcej innowacji jest w tym przypadku wprowadzanych dla napędów hybrydowych, a na drugim miejscu są napędy elektryczne. Liczba innowacji związanych z silnikami spalinowymi odzwierciedlana liczbą zgłoszeń patentowych jest więc 10x mniejsza niż w przypadku silników hybrydowych.

Przeszukajmy teraz Google Patents z wykorzystaniem poniższej frazy "((electric) OR (gas) OR (oil) OR (hydrogen)) (engine)". Użycie spójnika OR powoduje, że zapytanie staje się bardzo szerokie i w istocie liczba wyników jest dość znaczna – 4.827.461. Analizie takie liczby zgłoszeń

byłaby zbyt czasochłonna dlatego należałoby zawęzić zapytanie. Dla oceny stanu badań na rynku takie ogólne zagregowane dane mogą być jednak interesujące. Na przykład Google Patents pokazuje statystyki dotyczące wynalazców, przedstawione na rys **6.32**. Można z niego wyczytać, że o ile w latach 70. liderem innowacji był Nissan, to od roku 2004 najwięcej zgłoszeń patentowych ma Toyota, a jej ogólny udział w zgłoszeniach wynosi 8,2%.

Gdybyśmy chcieli teraz sprawdzić, w jakiej kategorii Toyota miała najwięcej zgłoszeń, to sam interfejs Google Patents tego nie umożliwia. Można to jednak uzyskać dzięki bezpośredniego dostępowi do danych patentowych, które Google udostępnił jako Google Patents Public Datasets. Łącznie dostępne jest 21 zbiorów (rys. **6.33**).

RYSUNEK 6.31. **KLASA B60 W KLASYFIKACJI ESPACENET**



TRANSPORTING

B60	**VEHICLES IN GENERAL**
B60K	**ARRANGEMENT OR MOUNTING OF PROPULSION UNITS OR OF TRANSMISSIONS IN VEHICLES; ARRANGEMENT OR MOUNTING OF PLURAL DIVERSE PRIME-MOVERS IN VEHICLES; AUXILIARY DRIVES FOR VEHICLES; INSTRUMENTATION OR DASHBOARDS FOR VEHICLES; ARRANGEMENTS IN CONNECTION WITH COOLING, AIR INTAKE, GAS EXHAUST OR FUEL SUPPLY OF PROPULSION UNITS IN VEHICLES**
	Arrangement or mounting of propulsion units in vehicles (of control devices for such units B60K 26/00; elastic mountings per se F16F; propulsion units or their control per se, see the relevant classes)
B60K 1/00	**Arrangement or mounting of electrical propulsion units** (B60K 7/00 takes precedence; arrangement or mounting of plural diverse prime-movers for mutual or common propulsion B60K 6/00; electric transmission arrangements B60K 17/12; electric equipment or propulsion of electrically-propelled vehicles per se B60L; current-collectors for power supply lines of electrically-propelled vehicles B60L 5/00)
B60K 3/00	**Arrangement or mounting of steam or gaseous-pressure propulsion units** (B60K 7/00 takes precedence; arrangement or mounting of plural diverse prime-movers for mutual or common propulsion B60K 6/00; gaseous-pressure transmission arrangements B60K 17/10)
B60K 5/00	**Arrangement or mounting of internal-combustion or jet-propulsion units** (B60K 7/00 takes precedence; arrangement or mounting of plural diverse prime-movers for mutual or common propulsion B60K 6/00)
B60K 6/00	**Arrangement or mounting of plural diverse prime-movers for mutual or common propulsion, e.g. hybrid propulsion systems comprising electric motors and internal combustion engines** {; Control systems therefor, i.e. systems controlling two or more prime movers, or controlling one of these prime movers and any of the transmission, drive or drive units (arrangement or mounting in vehicles of electrical gearing, in which an electrical machine serves only as reduction gearing and not as the prime mover and in which no electrical storing means are used B60K 17/12; control and regulation of purely electrical prime movers B60L; prime-movers comprising electrical and internal combustion motors in a common engine block or housing per se F02B 65/00; electric motors or motor-generators used for starting the combustion engine F02N 11/04; electric motors for synchronising gearing F16H 3/12)}{Informative references: mechanical gearings with secondary electric drive F16H 3/72; arrangements for handling mechanical energy structurally associated with the dynamo-electric machine H02K 7/00; machines comprising structurally interrelated motor and generator parts H02K 51/00; dynamo-electric machines not otherwise provided for in H02K see H02K 99/00}
B60K 7/00	**Disposition of motor in, or adjacent to, traction wheel** (roller-skate driving mechanisms A63C 17/12)
B60K 8/00	**Arrangement or mounting of propulsion units not provided for in one of the preceding main groups**

Źródło: https://pl.espacenet.com/classification?locale=pl_PL#!/CPC=B60K

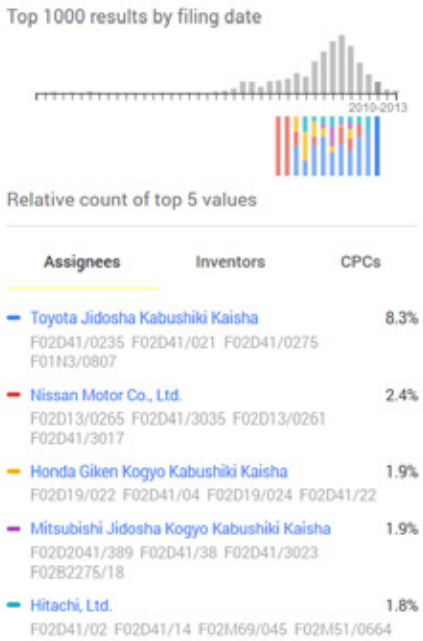
Najciekawsze są 'Google Patents Public Data' oraz 'Google Patents Research'. Dostęp do tych danych odbywa się poprzez narzędzie BigQuery. Na rys. **6.34** została przedstawiona struktura tabeli jednej z kluczowych tabel, tj. patents.publications. Samo zapytanie można sformułować w języku SQL.

TABELA 6.1. **LICZBA ZGŁOSZEŃ PATENTOWYCH W KATEGORII B60K W GOOGLE PATENTS**

KLASA	OPIS	LICZBA WYNIKÓW
B60K1	Napędy elektryczne	50.771
B60K3	Napędy parowe i ciśnieniowe	366
B60K5	Napędy wewnętrznego spalania	15.409
B60K6	Napędy hybrydowe	148.797
B60K8	Inne napędy	0

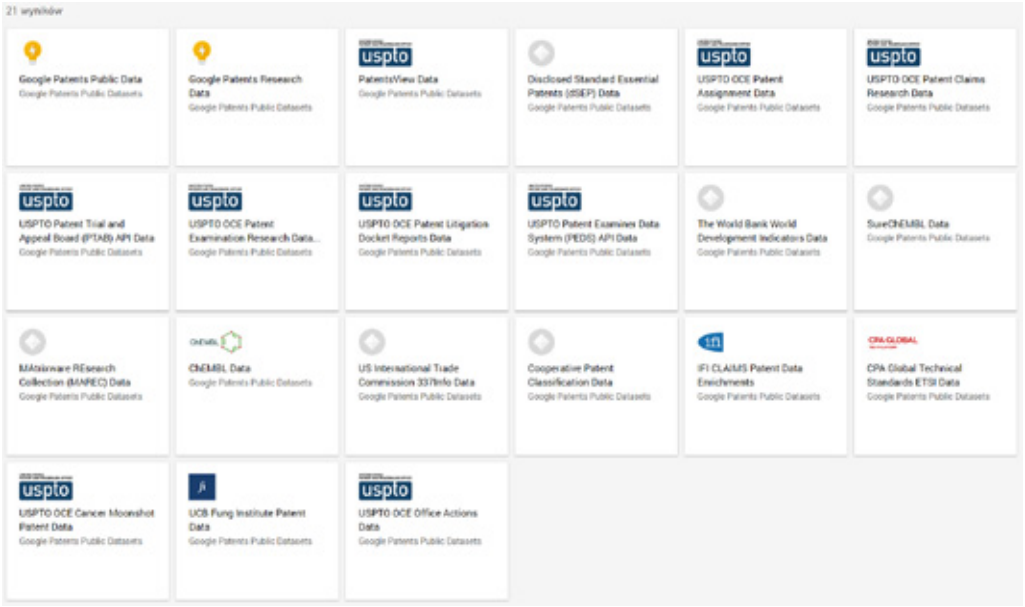
Źródło: Opracowanie własne

RYSUNEK 6.32. **WYNALAZCY W ZAKRESIE NAPĘDÓW DO SAMOCHODÓW WG LICZBY ZGŁOSZEŃ PATENTOWYCH W GOOGLE PATENTS**



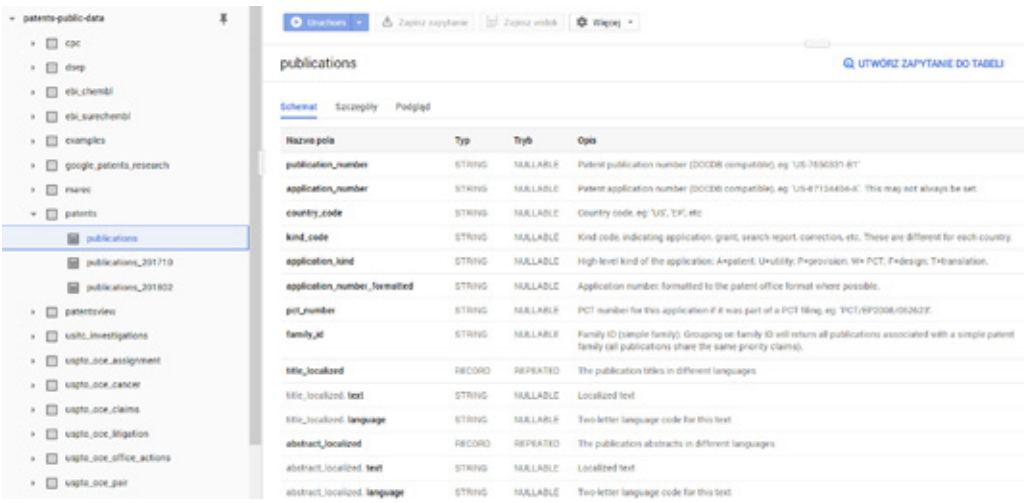
Źródło: https://patents.google.com/?q=electric,gas,oil,hydrogen&q=engine

RYSUNEK 6.33. **ZBIORY GOOGLE PATENTS PUBLIC DATASETS**



Źródło: https://console.cloud.google.com/marketplace/partners/patentspublic-data

RYSUNEK 6.34. **INTERFEJS BIGQUERY DO DANYCH PUBLICZNYCH GOOGLE PATENTS**



Źródło: zrzut ekranu.

6.6.2. Uzyskanie danych dotyczących energetyki wiatrowej

Przedsiębiorca jest zainteresowany pozyskaniem danych dotyczących wiatru. Załóżmy, że chcemy przygotować szybką demonstrację w Pythonie, dlatego skorzystamy ze źródła OpenML (zob. sekcja **6.5.2**). W portalu OpenML wpisujemy term do wyszukanie 'wind'. Otrzymujemy kilkanaście wyników. Decydujemy się na zbiór o nazwie wind, który zgodnie z opisem zawiera średnie dzienne prędkości wiatrów wyrażone w węzłach zmierzone w latach 1961-1978 w 12 stacjach meteorologicznych w Irlandii. W adresie URL odnajdujemy identyfikator zbioru, który będzie potrzebny do jego pobrania w przygotowywanym skrypcie – tutaj 503.

Istotną zaletą OpenML jest integracja z Pythonem. Możemy napisać skrypt bezpośrednio w Pythonie, jednak tutaj dla celów demonstracyjnych skorzystamy ze środowiska Jupyter[45]. W pierwszej komórce wpisujemy kod związany z wykorzystywanymi modułami oraz konfigurujemy wyświetlanie wykresów bezpośrednio w przeglądarce.

```
%matplotlib inline
from openml import tasks, flows, datasets
import pandas as pd
import matplotlib.pyplot as plt
```

W kolejnej komórce wczytujemy zbiór. Należy zwrócić uwagę, że należy podać identyfikator zbioru określony z adresu URL – 503.

```
ds = datasets.get_dataset(503)
X, cols = ds.get_data(return_attribute_names=True)
df = pd.DataFrame(X, columns=cols)
df
```

	year	month	day	RPT	VAL	ROS	KIL	SHA	BIR	DUB	CLA	
0	61.0	1.0	1.0	15.040000	14.960000	13.170000	9.290000	13.960000	9.870000	13.670000	10.250000	10.8
1	61.0	1.0	2.0	14.710000	16.879999	10.830000	6.500000	12.620000	7.670000	11.500000	10.040000	9.79
2	61.0	1.0	3.0	18.500000	16.879999	12.330000	10.130000	11.170000	6.170000	11.250000	8.040000	8.50
3	61.0	1.0	4.0	10.580000	6.630000	11.750000	4.580000	4.540000	2.880000	8.630000	1.790000	5.83
4	61.0	1.0	5.0	13.330000	13.250000	11.420000	6.170000	10.710000	8.210000	11.920000	6.540000	10.9
5	61.0	1.0	6.0	13.210000	8.120000	9.960000	6.670000	5.370000	4.500000	10.670000	4.420000	7.17

45 https://jupyter.org/

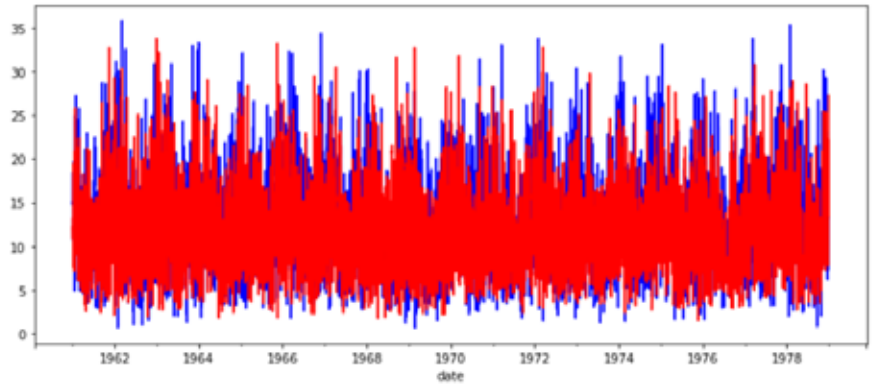
```
df['d'] = df.year*10000 + df.month*100 + df.day +  19000000
df['date'] = pd.to_datetime(df.d,format='%Y%m%d', box=True)
df.set_index(df.date, inplace=True)
df
```

	year	month	day	RPT	VAL	ROS	KIL	SHA	BIR	DUB	CL
date											
1961-01-01	61.0	1.0	1.0	15.040000	14.960000	13.170000	9.290000	13.960000	9.870000	13.670000	10.25000
1961-01-02	61.0	1.0	2.0	14.710000	16.879999	10.830000	6.500000	12.620000	7.670000	11.500000	10.04000
1961-01-03	61.0	1.0	3.0	18.500000	16.879999	12.330000	10.130000	11.170000	6.170000	11.250000	8.040000
1961-01-04	61.0	1.0	4.0	10.580000	6.630000	11.750000	4.580000	4.540000	2.880000	8.630000	1.790000
1961-01-05	61.0	1.0	5.0	13.330000	13.250000	11.420000	6.170000	10.710000	8.210000	11.920000	6.540000

W naszym zbiorze data jest rozbita na poszczególne elementy. Teraz łączymy dzień, miesiąc oraz rok i za pomocą funkcji to_datetime() zamieniamy te pola na jedno pole indeksu.

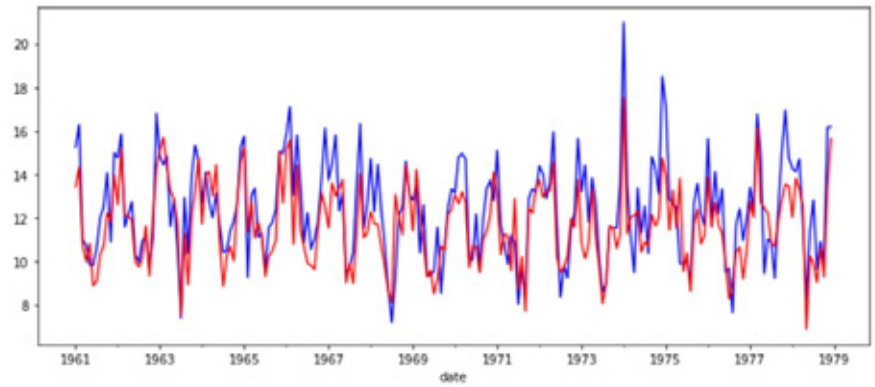
W kolejnym kroku przygotowujemy wykres dla dwóch wybranych lokalizacji (RPT oraz POS).

```
plt.figure(figsize=(12,5))
ax1 = df.RPT.plot(color='blue')
ax2 = df.ROS.plot(color='red')
plt.show()
```



Okazuje się, że jest on mało czytelny, gdyż mamy do czynienia z danymi dziennymi. Dlatego też w kolejnym kroku dokonujemy agregacji wg miesięcy, obliczając średnią i pokazujemy dane zagregowane na wykresie.

```
dfg = df.groupby(pd.Grouper(key='date', freq='1M')).mean()
plt.figure(figsize=(12,5))
plt.xlabel('Lata')
ax1 = dfg.RPT.plot(color='blue')
ax2 = dfg.ROS.plot(color='red')
plt.show()
```



6.7. Podsumowanie

Analiza dostępności zbiorów danych w obszarze open science wskazuje, że środowisko naukowe może być dobrym źródłem danych. Nie powinniśmy tutaj przyjmować uogólnienia, że to uczelnie są dostawcą danych. To raczej poszczególni naukowcy lub grupy badawcze wychodzą z inicjatywą udostępnienia materiałów z badań. Często są one wynikiem projektu, gdy było zapewnione odpowiednie dofinansowanie na uporządkowanie zbiorów. W pewnej części jest to determinowane wymogami instytucji finansujących naukę, które z kolei działają w ramach określanych przez takie inicjatywy, jak OpenAIRE.

W poszukiwaniu zbiorów danych o wysokiej jakości warto zwracać uwagę na inicjatywy oddolne. Jednym z najlepszych przykładów jest kolekcja zbiorów danych o uczeniu maszynowym (zob. sekcja **6.5.1**). Istotną zaletą korzystania ze zbiorów danych udostępnionych w ramach open science jest to, że dane są powiązane z konkretnym artykułem, który nie tylko wyjaśnia dane, ale również proponuje metody ich analizy.

Najbardziej wartościowe zbiory danych to OpenML oraz Google Patents. OpenML umożliwia bezpośrednie korzystanie ze zbiorów do uczenia maszynowego w języku Python. Są tam zawarte zarówno klasyczne zbiory, wykorzystywane w nauce machine learning, jak i nowe zbiory pozwalające na tworzenie nowych usług. Google Patents to nie tylko przeszukiwalna baza patentów z najważniejszych jurysdykcji, ale przede wszystkim baza danych, którą można odpytać za pomocą języka SQL. Stwarza to zupełnie nowe możliwości w zakresie prowadzenia analiz stanu techniki.

Do poszukiwania zbiorów danych udostępnionych w ramach open science najlepiej wykorzystać dedykowane repozytoria. Największe z nich to Harvard Dataverse oraz Zenodo. Ta pierwsza baza została stworzona przez jedną z najlepszych uczelni na świecie. Co ciekawe, nie tylko umożliwiają deponowanie zbiorów na własnej infrastrukturze, ale również udostępniają oprogramowanie, dzięki czemu zainteresowane organizacje mogą tworzyć własne repozytoria danych.

7 Zagadnienia jakości informacji

7.1. Definicja

Jakość udostępnionych danych jest jednym z ważnych czynników, które decydują o tym, w jakim stopniu mogą one zostać ponownie wykorzystane w celu uzyskania wartości dodanej [**131**]. Można ją oceniać na podstawie wielu kryteriów, a sama ocena może dotyczyć zarówno danych, jak również metadanych je opisujących.

Bardzo ważną cechą jakości danych jest adekwatność do wymagań użytkownika końcowego [**18**]. Dane wysokiej jakości muszą spełniać określone kryteria, które są definiowane w zależności od standardów. Jednym z nich jest ISO 8000. Zgodnie z tą normą, w celu zapewnienia odpowiedniej jakości danych, należy uwzględnić szereg kryteriów, takich jak: dokładność, źródło, kompletność, cel (przeznaczenie) danych, metoda pomiaru lub oszacowania [**64**]. Ten standard opisuje również działania niezbędne do sprawdzenia wskaźników jakości danych [**78**], na przykład:

- sprawdzenie zgodności danych ręcznie lub przy użyciu narzędzia,
- statystyczna analiza wskaźników jakości danych, lub
- mogą być brane pod uwagę otwarte słowniki techniczne zgodnie z ISO/TS 2274530:2009.

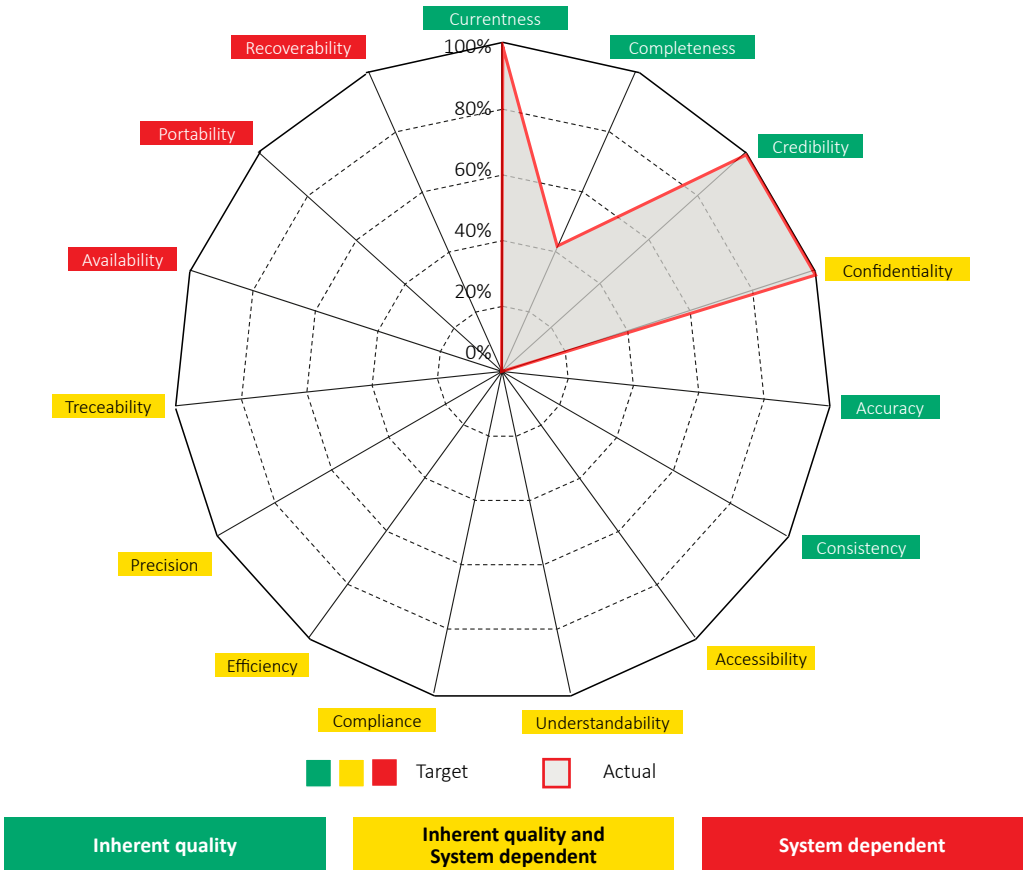
Innym kluczowym standardem jakości danych jest ISO/IEC 25012 [**5**, **126**]. Rys. **7.1** przedstawia kryteria jakości danych w oparciu o ten standard.

W literaturze nie ma wśród badaczy zgodności co do sposobu rozróżnienia dwóch pojęć: jakości danych i jakości informacji. Mandrick i inni zaznaczyli, że istnieje tendencja do używania terminu jakości danych w odniesieniu do zagadnień technicznych (np. integracja danych), natomiast jakość informacji – do zagadnień nietechnicznych, np. relewancji dla konkretnego odbiorcy [**113**]. Z drugiej strony, oba pojęcia można scharakteryzować jako rozbieżność pomiędzy wizją świata dostarczoną przez system informacyjny i prawdziwego stanu świata [**139**].

Metody i kryteria oceny jakości informacji dobierane są w zależności od rodzaju informacji, wśród których wyróżnić można ekonomiczną, medyczną, techniczną, itp. Na przykład Komisja Wspólnot Europejskich przyjęła specjalne kryteria oceny jakości dla stron internetowych związanych z ochroną zdrowia – jakość mierzy się na podstawie: przejrzystości, uczciwości, dostępności, wiarygodności, aktualności, ochrony danych osobowych, odpowiedzialności [**33**].

Pomiaru jakości informacji zazwyczaj dokonuje się w odniesieniu do pewnej równowagi społecznej, aby ustalić, co stanowi dobrą jakość informacji w danym kontekście aktywności [**169**]. Sam kontekst składa się z dwóch głównych elementów: kultury (języka, normy) i struktur społeczno-technicznych (w tym stosunków gospodarczych i standardów).

RYSUNEK 7.1. **JAKOŚĆ DANYCH W OPARCIU O ISO 25012**



Źródło: [126]

7.2. Wymiary jakości

Określenie wymiarów jakości informacji zależy od miejsca, w którym ta informacja została umieszczona, np. w drukowanej książce czy na portalu internetowym. Dla takiego źródła jak encyklopedia, w jego tradycyjnym rozumieniu, zdefiniowano 7 wymiarów jakości [**36**]: format, unikatowość, autorytet, zakres, dokładność, obiektywność oraz aktualność. Dodatkowo, zestaw wymiarów jakości informacji może się różnić w zależności od typu ocenianej informacji – np. mapy geograficzne, obrazy, teksty [**16**].

Dodatkowym problemem jest brak ogólnie przyjętego zestawu wymiarów jakości – istnieją publikacje, które takich wymiarów definiują ponad 40 [**50**, **153**]. Jednocześnie należy brać

pod uwagę, że takie obszerne listy wymiarów muszą być znacząco skrócone, ponieważ zawierają niespójności i powtórzenia [**50**]. Niektóre wymiary są pochodną innych wymiarów – do takich należą np. wiarygodność, autorytet lub reputacja, ponieważ mogą być wynikiem analizy poprawności i spójności informacji. Ponadto wymiary muszą być przydatne do użycia w praktyce, stąd niektórzy badacze posługują się tylko 6-8 wymiarami przy ocenie jakości danych oraz informacji [**16**, **36**, **87**].

7.2.1. Wikipedia

W celu zdefiniowania wymiarów jakości stron serwisów typu wiki, należy wziąć pod uwagę ich podobieństwo z tradycyjnymi encyklopediami oraz dokumentami Web 2.0. Większość dotychczasowych badań w zakresie jakości informacji w serwisach wiki skupiała na najbardziej popularnym przedstawicielu tychże serwisów – Wikipedii. Wyniki badań pokazują, że treść współtworzona przez użytkowników w Wikipedii może być uznana za encyklopedyczną, ponieważ ma taką samą dokładność, jak w przypadku tradycyjnych encyklopedii [**59**].

Jakość artykułu w tradycyjnej encyklopedii może być definiowana w 7 ogólnych wymiarach [**36**]:

- **format** – na jakim nośniku jest utrwalona informacja,
- **unikatowość** – cechy, które odróżniają ją od innych encyklopedii,
- **autorytet** – reputacja osób, które sprawdzają informację (recenzenci),
- **zakres** – jakie tematyczne obszary pokrywa encyklopedia, do jakiego użytkownika skierowana, w jakim stylu napisana,
- **dokładność** – czy zawiera błędy,
- **obiektywność** – wszechstronność i neutralność materiału, uwzględnienie ilustracji i innych materiałów multimedialnych,
- **aktualność** – zgodność z rzeczywistością w momencie, gdy jest użytkowana.

Powyższe kryteria częściowo pokrywają się z kryteriami ustalonymi przez społeczność Wikipedii. Autorytet w mniejszym stopniu dotyczy Wikipedii – artykuły nie muszą być w niej sprawdzane przez ekspertów. Natomiast w Wikipedii istotne są następujące kryteria: neutralny punkt widzenia, powoływanie się na oryginalne badania zamiast ich prezentacji, weryfikowalność.

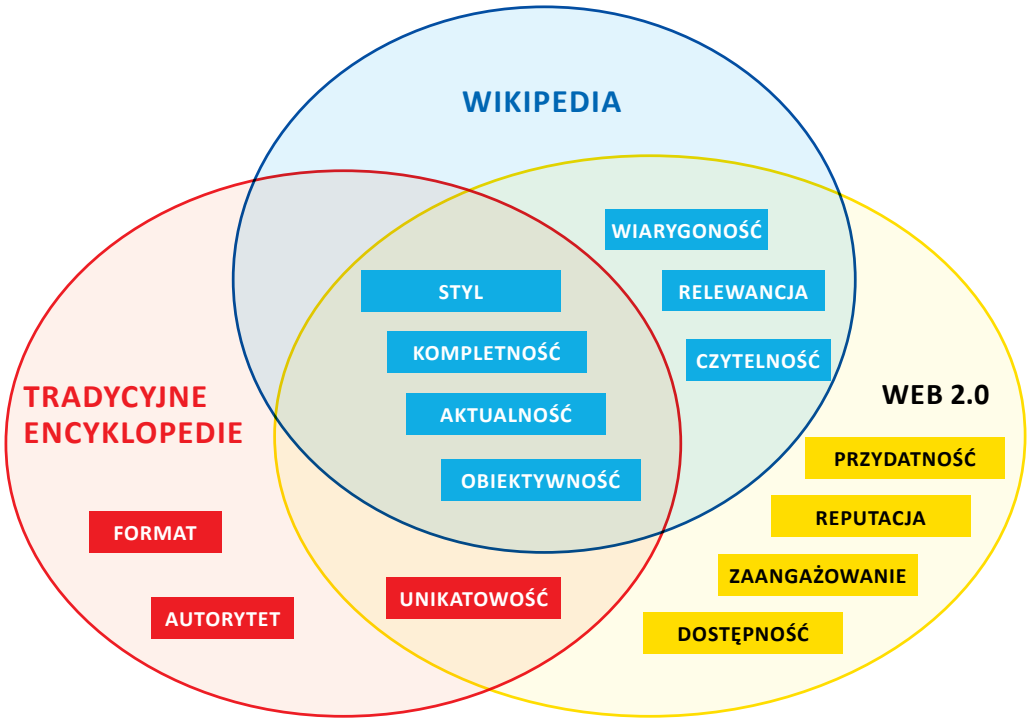
Chociaż Wikipedia funkcjonuje jako encyklopedia, to jednak istotnie różni się od tradycyjnego (drukowanego) odpowiednika – artykuły w Wikipedii mogą być tworzone oraz redagowane w czasie rzeczywistym przez użytkowników (także anonimowych), a zmiany są widoczne od razu dla czytelników.

Serwisy wiki działają na zasadach Web 2.0 [**45**], gdzie najważniejszą cechą jest to, że treści mogą być generowane przez użytkowników Internetu [**42**]. W konsekwencji pojawiły się nowe

rodzaje repozytoriów wiedzy, do współtworzenia których każdy może się swobodnie przyczynić. Przykładami takich repozytoriów mogą być serwisy oparte na mechanizmie pytania – odpowiedzi (Q&A), cyfrowe repozytorium wideo, blogi, witryny z recenzjami, serwisy społecznościowe oraz inne. Na podstawie literatury [**42**, **70**, **123**, **147**, **153**] oraz własnych obserwacji dla oceny jakości dokumentów Web 2.0 zostały zdefiniowane następujące wymiary: unikatowość, styl, kompletność, aktualność, obiektywność, weryfikowalność, relewancja, czytelność, przydatność, reputacja, zaangażowanie.

Pokrycie wymiarów jakości dla informacji z tradycyjnych encyklopedii, dokumentów Web 2.0 i Wikipedii pokazane jest na rys. **7.2**. Dla każdego z wymiarów istnieje odpowiedni zbiór miar. Przykład jest pokazany w tabeli **7.1**.

RYSUNEK 7.2. **WYMIARY JAKOŚCI DLA INFORMACJI Z RÓŻNYCH OBSZARÓW**



Źródło: [97]

TABELA 7.1. **WYMIARY JAKOŚCI SERWISÓW WIKI**

WYMIAR	OPIS	SPOSÓB MIERZENIA
Aktualność	Na ile artykuł opisuje obecny stan pewnej rzeczywistości.	Data ostatniej modyfikacji artykułu, częstość wprowadzania zmian.
Kompletność	Na ile wyczerpujący jest opis tematu	Długość artykułu i poszczególnych jego elementów (abstraktu, tabeli, sekcji…).
Weryfikowalność	Na ile przedstawione informacje można sprawdzić.	Liczba referencji, jakość referencji, wskaźnik ref/długość, liczba referencji na sekcję, czy zawiera specjalne szablony wskazujące na problemy z weryfikowalnością.
Styl	W jaki sposób jest organizowana zawartość artykułu	Liczba i kolejność sekcji, liczba tabel, liczba linków, liczba szablonów, liczba obrazków, czy zawiera specjalne szablony wskazujące na luki w stylu.
Obiektywność	Na ile zawartość artykułu spełnia kryterium neutralnego punktu widzenia, czy zawiera obrazki i inne materiały multimedialne dotyczące tego artykułu.	Liczba obrazków i innych elementów multimedialnych dotyczących tego artykułu (np. nie liczone są obrazki ze znaczkami flag), specjalne szablony, analiza NLP.
Czytelność	Na ile tekst jest zrozumiały i wolny od niepotrzebnej złożoności.	Wskaźnik Flesch-Kincaid, wskaźniki użycia różnych słów (w tym fachowych).
Relewancja	Na ile artykuł jest ważny dla użytkowników.	Liczba wyświetleń, liczba obserwujących użytkowników, ocena ważności przez Wikiprojekty, liczba linków na artykuł, PageRank, liczba autorów.

Źródło: opracowanie własne.

7.2.2. Linked Open Data

Termin Linked Open Data (LOD) odnosi się do powiązanych ze sobą i semantycznie opisanych zbiorów danych strukturalnych. Zbiory te wykorzystują standardowe technologie internetowe, takie jak HTTP[1], RDF[2] oraz URI[3], które umożliwiają automatyczne odczytywanie oraz zrozumienie danych przez komputery.

LOD służy do budowania semantycznie opisanych baz wiedzy, jak np. Wikidane oraz DBpedia, które ze względu na rozmiar tworzone są przez społeczność. Dla takich baz można wyróżnić następujące wymiary jakości [**97**, **143**, **202**]:

- *Precyzja* – stopień, w jakim dane są akceptowane jako prawdziwe i wolne od błędów.
- *Obiektywność* – stopień, w jakim dane są wolne od uprzedzeń i bezstronności.
- *Wiarygodność* – czy źródła określone dla danych są godne zaufania.

1 Hypertext Transfer Protocol
2 Resource Description Framework
3 Uniform Resource Identifier

- *Spójność* – stopień, w jakim dane są zgodne lub przyczyniają się do tworzenia spójnej reprezentacji wiedzy.
- *Aktualność* – czy dane mają bieżące wartości zgodne z rzeczywistością.
- *Kompletność* – czy dane mają wystarczającą szerokość, głębokość i zakres. Tu można wyróżnić trzy rodzaje kompletności: schematu, pozycji, populacji.
- *Interpretacja* – zakres, w jakim dane mogą być jednoznacznie interpretowane przez maszyny.
- *Zrozumiałość* – stopień, w jakim dane mogą być jednoznacznie zinterpretowane przez człowieka.
- *Powiązanie* – stopień, w jakim dane są odnoszą się do innych zasobów.

Dane do projektu Wikidane trafiają na trzy sposoby. Po pierwsze, użytkownicy dodają indywidualne fakty do bazy (najpopularniejszy sposób). Po drugie, Wikidane wykorzystuje dane z innych projektów Wikimedia, które jednak są czasami nieaktualne, niekompletne lub niepoprawne. Trzeci sposób to import dużych zewnętrznych zestawów danych.

Społeczność Wikidanych zdefiniowała klasy jakości dla elementów tej bazy wiedzy [**192**] (od najwyższej):

- Klasa A – wszystkie dostępne informacje są wprowadzone z wiarygodnymi referencjami.
- Klasa B – element zawiera wszystkie najważniejsze stwierdzenia, z dobrymi referencjami, tłumaczeniami, aliasami, linkami do podstron oraz obrazów.
- Klasa C – dostępna jest większość podstawowych informacji, których można się spodziewać. Mogą one nie być dobrze udokumentowane referencjami lub być niekompletne.
- Klasa D – zawiera wystarczającą ilość informacji, aby łatwo zidentyfikować element.
- Klasa E – wszystko, co nie spełnia wymagań klas A-D.

Szczególną uwagę należy zwrócić na elementy Wikidanych posiadające najwyższą klasę A, które muszą zawierać m.in.:

- wiele zewnętrznych odniesień (referencji) do nietrywialnych stwierdzeń;
- tłumaczenia wykonane dla najistotniejszych języków (dotyczy to etykiet oraz opisów);
- wszystkie linki do odpowiednich stron, które istnieją w innych serwisach wiki;
- wszystkie odpowiednie aliasy (przekierowania czy inne nazwy), które istnieją w najważniejszych językach;
- obraz wysokiej jakości (jeżeli dotyczy).

Przy ocenie kompletności elementów na określony temat brane są pod uwagę odpowiednie stwierdzenia. Na przykład element opisujący człowieka (identyfikator Q5) powinien zawierać stwierdzenia z właściwościami płci, datą urodzenia, miejscem urodzenia[4]. Zaleca się stosowanie

4 Przykład elementu Wikidanych opisujący człowieka: https://www.wikidata.org/wiki/Q935

własnych kryteriów kompletności pozycji w celu oceny, czy każda pozycja zawiera wszystkie odpowiednie stwierdzenia. Elementy Wikidanych o najwyższej jakości muszą zawierać referencje do co najmniej dwóch niezależnych publikacji.

7.2.3. Zbiory danych

Jedną z przesłanek do oceny jakości zbiorów danych może być ich popularność. Można przypuszczać, że im częściej użytkownicy używają określonego zbioru danych, tym jego jakość jest wyższa. Grafy wiedzy ogólnego przeznaczenia, takie jak DBpedia, YAGO oraz Wikidata, stały się centralną częścią chmury danych LOD [**154**] i należą do najczęściej używanych zestawów danych w sieci [**48**]. Takie bazy wiedzy zawierają informacje na temat milionów obiektów z różnych obszarów.

Zbiory danych częściej podlegają ponownemy wykorzystaniu, jeżeli są powiązane z innymi zbiorami danych, przytaczane w publikacjach lub dyskusjach w serwisach społecznościowych. Można zatem określić wartość miary ponownego wykorzystania danych, podobnie jak robi się to w odniesieniu do cytowań artykułów [**48**].

Jednym z najbardziej rozpowszechnionych otwartych zbiorów danych jest Open Government Data (OGD), który odnosi się do informacji gromadzonych przez organy administracji publicznej i udostępnionych w Internecie. Istnieją różne podejścia do oceny takich baz danych. Na przykład OpenDataMonitor[5] udostępnia kompleksowy zestaw wskaźników oceny jakości otwartych źródeł danych/portali, takie jak: otwarte licencje, możliwosc automatycznego przetwarzania przez maszyny, dostępność oraz kompletność metadanych. Inny przykład to projekt COMSODE[6], który wprowadza koncepcję platformy otwartych danych, wspierającą ponowne wykorzystanie otwartych danych poprzez ich wzbogacanie i łączenie. Kolejnym przykładem jest stosunkowo nowa inicjatywa ADEQUATe, która zamierza aktywnie angażować wszystkie strony związane z otwartymi danymi, aby przyczynić się do ich tworzenia, ulepszania, wzbogacania, przeglądania, ponownego wykorzystywania i rozpowszechniania [**174**].

Miary jakości muszą opierać się na następujących zasadach [**69**]:

- *Mierzalność* – miary muszą być normalizowane, a co najmniej skalowalne.
- *Interpretacja* – miary muszą być zrozumiałe. Ich definicja powinna zawierać odpowiednią ilość informacji, aby można było je poprawnie zinterpretować.

Agregacja – istnieje możliwość ilościowego określenia jakości danych na poziomie atrybutów, a także na poziomie rekordu, zbioru danych lub bazy danych. W ten sposób miary będą miały

5 https://www.opendatamonitor.eu
6 https://www.comsode.eu

semantyczną spójność na wszystkich poziomach. Ponadto miary powinny zezwalać na agregację wartości, aby uzyskać miarę na wyższym poziomie.

- *Wykonalność* – żeby miary mogły być stosowane w praktyce, powinny być oparte na możliwych do określenia parametrach wejściowych, a proces zbierania i liczenia powinien być w miarę możliwości zautomatyzowany.

Miary można sklasyfikować jako obiektywne, gdy są oparte na pomiarach ilościowych, lub subiektywne, gdy są oparte na ocenach jakościowych pochodzących od użytkowników.

Dla otwartych zbiorów danych możemy zdefiniować następujące wymiary jakości [**180**]:

- *Kompletność* – na ile dane są pełne, czy zawierają określone metadane.
- *Czytelność* – czy format jest zgodny z dobrze znanym standardem.
- *Identyfikowalność* – czy jest możliwość sprawdzenia źródła danych.
- *Spójność* – w jakim stopniu dane są niespójne.
- *Aktualność* – na ile dane pokazują stan rzeczywisty.
- *Dokładność* – w jakim stopniu baza zawiera błędy.
- *Zrozumiałość* – ile czasu trzeba poświęcić na zrozumienie danych.

Ponadto do oceny jakości mogą być przydatne miary określające wolumen danych: liczba rekordów i objętość. Tabela **8.4** przedstawia listę miar wraz z ich opisem.

7.3. Mierzenie jakości źródeł

Wybór odpowiedniego zbioru danych może zależeć od źródła, w którym jest on dostępny lub opisany. Do takich źródeł mogą należeć również prace naukowe.

7.3.1. Mierzenie jakości tekstu

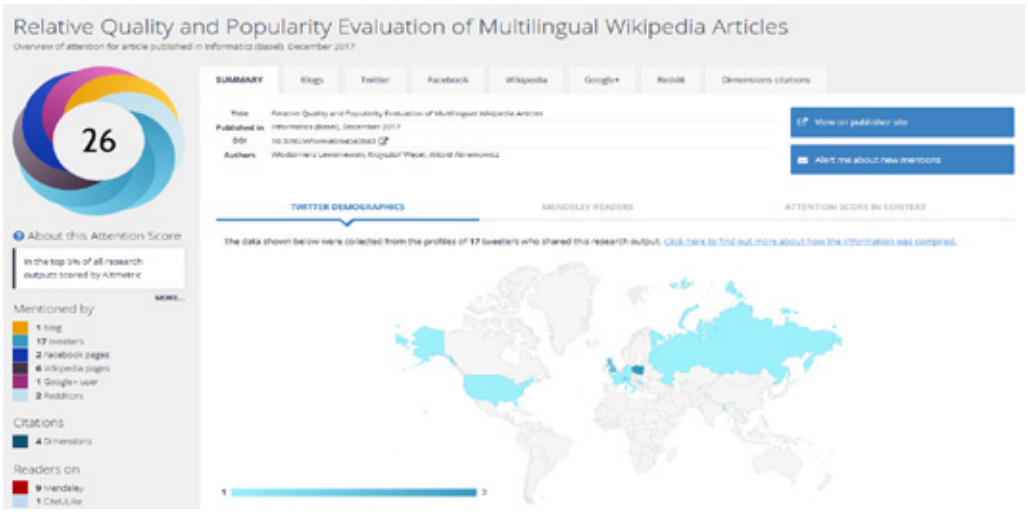
Badania akademickie muszą być oceniane na wielu różnych poziomach, aby wspierać proces decyzyjny. Liczba cytowań jest jedną z najpopularniejszych miar oceny jakości publikacji naukowych. Podstawową przyczyną, dla której liczba cytowań może odzwierciedlać wpływ naukowy, jest fakt, że odniesienia w dokumencie wymieniają wcześniejsze prace, które w pewnym stopniu miały na niego wpływ. Dokument, który został zacytowany w wielu różnych publikacjach, miał zatem bardzo duży wpływ. Na tej podstawie porównanie liczby cytowań dokumentów lub grup dokumentów może ujawnić, który był najbardziej istotny. Pomiar liczby cytowań ma jednak pewne wady:

- Potrzeba około trzech lat, aby liczba cytowań była wystarczająco dojrzała i była rozsądnym wskaźnikiem długotrwałego oddziaływania artykułu naukowego [**61**].

- Badania mogą być przydatne w środowisku akademickim w sposób, który nie generuje cytowań, np. zamykanie fałszywych obszarów badań lub przekazywanie informacji pracownikom akademickim w sposób, który nie jest konwencjonalny [**157**].
- Decyzja o tym, które prace cytować, może wynikać z uwarunkowań takich, jak narodowość autora, prestiż, uniwersytet i współpraca [**140**].
- Średnia liczba cytowań różni się w zależności od rodzaju dokumentu, obszaru badań i roku, a więc nie jest sprawiedliwe porównywanie liczby cytowań między dokumentami z różnych dziedzin i lat [**7**].
- Niektóre cytaty są negatywne, pobieżne lub arbitralne [**93**].
- Cytaty mogą nie odzwierciedlać cennych nieakademickich zastosowań badań, więc mogą nie być dobrym wskaźnikiem wpływu społecznego [**177**].

W związku z powyższym poszukiwane są alternatywne miary, które mogą wspomóc obiektywną ocenę wpływu publikacji naukowych. Te alternatywne miary mogą mieć określone zalety, np. szybkość ich uzyskania lub zdolność do odzwierciedlenia nienaukowego lub społecznego wpływu badań [**173**]. Jednym z projektów, który mierzy jakość prac naukowych, jest Altmetric[7]. Miara jakości pracy naukowej określona przy pomocy tego narzędzia pokazana jest na rys. **7.3**.

RYSUNEK 7.3. **ALTMETRIC: PRZYKŁAD MIERZENIA JAKOŚCI PRACY NAUKOWEJ**



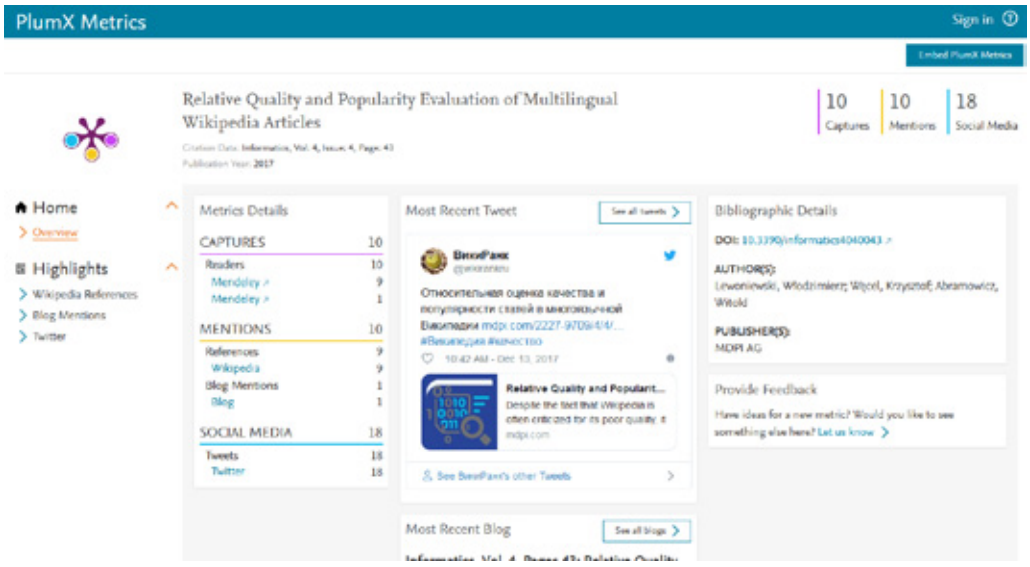
Źródło: https://mdpi.altmetric.com/details/30175469

7 https://www.altmetric.com/

Innym alternatywnym narzędziem do mierzenia jakości publikacji naukowych jest projekt PlumX[8]. Wynik mierzenia jakości pracy naukowej przy pomocy tego narzędzia pokazany jest na rys. **7.4**.

Często ta sama praca naukowa może być dostępna w różnych serwisach internetowych. W związku z tym, dobrze jest użyć specjalnych identyfikatorów, pomagających w znalezieniu tożsamych referencji, nawet w przypadku różnych parametrów w opisie (np. tytuły w innych językach). Można również ujednolicić ich adresy URL. Na przykład, jeśli numer referencyjny to ISBN „978-3-319-46254-7", adres URL będzie miał postać: books.google.com/books?vid=ISBN9783319462547. Bardziej szczegółowe informacje na temat identyfikatorów, które były wykorzystane do ujednolicenia odniesień, przedstawiono w tabeli **7.2**.

RYSUNEK 7.4. **PLUMX: PRZYKŁAD MIERZENIA JAKOŚCI PRACY NAUKOWEJ**



Źródło: https://plu.mx/plum/a/?doi=10.3390/informatics4040043

8 https://plu.mx

TABELA 7.2. **IDENTYFIKATORY UŻYWANE DO UNIFIKACJI ŹRÓDEŁ**

IDENT.	OPIS	NOWY URL
arXiv	arXiv repository identifier	http://arxiv.org/abs/...
DOI	Digital object identifier	http://doi.org/...
ISBN	International Standard Book Number	http://books.google.com/books?vid=ISBN...
ISSN	International Standard Serial Number	https://worldcat.org/ISSN/...
JSTOR	Journal Storage number	https://jstor.org/stable/...
PMC	PubMed Central	https://ncbi.nlm.nih.gov/pmc/articles/PMC...
PMID	PubMed	https://ncbi.nlm.nih.gov/pubmed/...
OCLC	WorldCat's Online Computer Library Center	https://worldcat.org/oclc/...

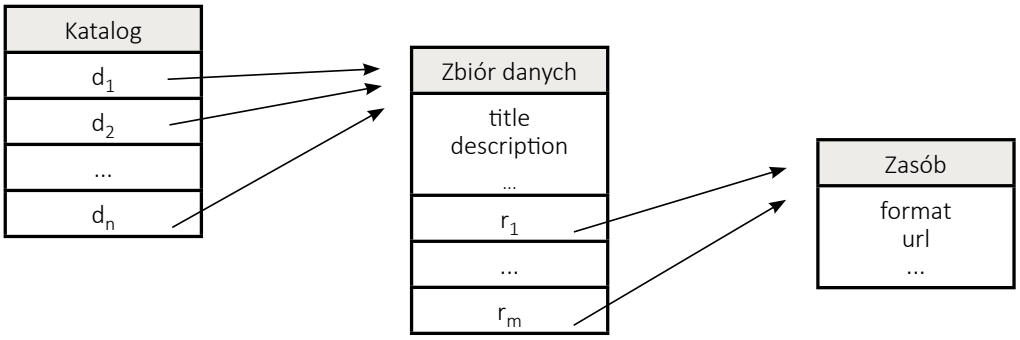
Źródło: [101]

7.3.2. Mierzenie jakości portali danych

Większość otwartych zbiorów danych jest publikowana na specjalnych portalach, które są w zasadzie katalogami podobnymi do bibliotek cyfrowych (patrz rys. **7.5**): zbiór danych agreguje grupę plików danych (nazywanych zasobami lub dystrybucjami), które są dostępne do odczytania lub pobraniu w jednym lub wielu formatach (np. CSV, PDF, Microsoft Excel [**131**]). Ponadto zbiór danych zawiera metadane (tj. podstawowe informacje opisowe w formacie strukturalnym) dotyczące tych zasobów, np. informacje o autorze, pochodzeniu lub licencji.

Serwisy internetowe, które umożliwiają umieszczanie zbiorów danych przez użytkowników, mogą dostarczać różne metainformacje, korzystne z punktu widzenia oceny jakości danych, np. informacje o popularności konkretnego zbioru danych.

RYSUNEK 7.5. **STRUKTURA WYSOKIEGO POZIOMU KATALOGU DANYCH**



Źródło: [131]

Otwarte dane najczęściej są tworzone i udostępniane przez organy administracji publicznej – Open Government Data (ODG). Jakość portali z otwartymi danymi można mierzyć biorąc po uwagę następujące kryteria:

- *Otwartość licencji* – na jakich zasadach możemy wykorzystywać dane umieszczone na portalu.
- Możliwość przetwarzania maszynowego – czytelność danych, która zależy od formatu.
- *Kompletność metadanych* – sprawdzenie, czy są zdefiniowane wszystkie (ważne) dane, opisujące zbiory danych: licencja, autor, organizacja, data aktualizacji danych oraz inne.
- *Dostępność*–udział procentowy zbiorów danychdostępnych publicznie w stosunku do całkowitej zawartości katalogu.
- *Popularność* – oszacowanie, jak ważny jest katalog w sieci, wykonywane w oparciu o Google, Alexa oraz inne systemy oceny popularności stron WWW.
- *Otwarte formaty* – udział procentowy danych dostępnych w formacie otwartym.
- *Zakres tematyczny* – czy portal dotyczy danych tylko o określonej tematyce (np. zbiory danych bibliograficznych, wskaźniki ekonomiczne).
- *Miary społecznościowe* – czy portal umożliwia zbieranie informacji zwrotnych, dotyczących poszczególnych zbiorów danych (np. liczba głosów, odwiedzin, użycia).

Poniżej został omówiony przykład analizy jakości portalu OpenML (zob. również sekcja **6.5.2**). Portal ten zawiera ponad 2500 zbiorów danych. Jest to miejsce, w którym naukowcy zajmujący się uczeniem maszynowym mogą automatycznie udostępniać dane i organizować je tak, aby ułatwić ich ponowne wykorzystanie. W odniesieniu do przedstawionych wyżej kryteriów portal OpenML może być oceniony następująco:

- Otwartość licencji – wszystkie zestawy danych mają licencje CC-BY (uznanie autorstwa). Dane mogą być swobodnie kopiowane, zmieniane, rozprowadzane, przedstawianie czy wykonywane, przy czym jedynym warunkiem jest poinformowanie o twórcy danych, źródle oraz samej licencji.
- Możliwość przetwarzania maszynowego – dostępny pakiet Python, który umożliwia bezpośrednie korzystanie z tych zbiorów we własnych programach.
- Kompletność metadanych – informacja o autorze, data aktualizacji, opis tekstowy, liczba instancji, typów danych, brakujących danych.
- Dostępność – wszystkie dane są publicznie dostępne.
- Popularność – według Alexa serwis nie jest obecnie zbyt popularny (ok. 290 tys. miejsce w rankingu najczęściej odwiedzanych stron WWW[9]). Jednak należy wziąć pod uwagę, że serwis ma ściśle określoną grupę docelową.

9 https://www.alexa.com/siteinfo/openml.org

- Otwarte formaty – dane dostępne w szerokiej gamie otwartych formatów: ARFF, CSV, JSON, XML, RDF.
- Zakres tematyczny – portal skierowany do naukowców, zajmujących się uczeniem maszynowym.
- Miary społecznościowe – każdy zestaw danych zawiera liczbę pobrań, polubień, wykorzystań, ocenę wpływu oraz zasięgu[10].

Podczas badania struktury metadanych różnych portali ze zbiorami danych można zaobserwować różne schematy metadanych. Na przykład, serwisy zbudowane na podstawie struktury Socrata opisują informacje licencyjne za pomocą pojedynczego parametru metadanych, podczas gdy portale wykorzystujące strukturę CKAN posiadają trzy różne parametry do określania ID, adresu URL oraz nazwy licencji.

W celu unifikacji różnych standardów opisów zbiorów danych można użyć modelu DCAT [**131**]. Na rys. **7.6** pokazano odwzorowanie parametrów metadanych z różnych struktur do modelu DCAT.

RYSUNEK 7.6. **MAPOWANIA PARAMETRÓW META DANYCH Z RÓŻNYCH STRUKTUR DO MODELU DCAT**

DCAT	CKAN	SOCRATA	OPENDATASOFT
dcat:Dataset			
→ dct:title	title	name	title
→ dct:description	notes	description	description
→ dct:issued	metadata_created	createdAt	-
→ dct:modified	metadata_modified	viewLastModified	modified
→ dct:identifier	id	id	datasetid
→ dcat:keyword	tags	tags	keyword
→ dct:language	language	-	language
→ dct:publisher	organization	owner	publisher
→ dct :contactPoint	maintainer, author (-email)	tableAuthor	-
→ dct :accrualPeriodicity	frequency	-	-
→ dct: landingPage	url	-	-
→ dct:theme	-	category	theme

Źródło: [131]

10 Więcej informacji odnośnie miar połecznych na stronie https://docs.openml.org/ altmetrics/

7.4. Analiza przypadków

W niniejszej sekcji zostały opisane różne aspekty badania oraz doskonalenia jakości źródeł danych. Motywem przewodnim jest Wikipedia, a analizowane są aspekty oceny jakości w Wikipedii oraz w projektach pochodnych, takich jak DBpedia. Przedstawiamy również narzędzia: OpenRefine do doskonalenia jakości zbiorów oraz Wikirank do oceny jakości artykułów Wikipedii.

7.4.1. Wikipedia

Wikipedia nie posiada centralnej redakcji czy grupy użytkowników, którzy mogliby metodycznie oceniać artykuły. Byłoby to bardzo trudne, ponieważ treść jest dynamiczna – może zmieniać się nawet kilka razy dziennie. Co więcej, autorzy nie muszą formalnie potwierdzać swoich umiejętności w określonej dziedzinie. Te i inne aspekty funkcjonowania Wikipedii były przedmiotem krytyki, w szczególności wskazującej na słabą jakość informacji.

Można się spodziewać, że osoby odwiedzające strony Wikipedii oraz jej twórcy są zainteresowani wysoką jakością treści w niej zawartej. Dotyczy to również dużych korporacji, które m.in. muszą dbać o obiektywne przestawienie informacji o swoich produktach. Na przykład w 2017 roku sieć restauracji Burger King w swojej reklamie proponowała wyszukać w wyszukiwarce Google informacje na temat ich nowego produktu, który był również opisany w Wikipedii. W tym czasie opis tego produktu w Wikipedii został zmieniony przez jednego z użytkowników – we wprowadzeniu zostało zaznaczone, że jest to „najgorszy produkt", a także dodane zostały inne niesprawdzone informacje [**183**].

Według Wikipedii jakość artykułu jest rozumiana jako jego zdolność do sprostania oczekiwaniom i potrzebom jego docelowych odbiorców [**194**].

Pojęcie jakości w Wikipedii nie jest uniwersalne w czasie i przestrzeni. Obecnie Wikipedia zawiera około 300 wersji językowych i każda ma swoją społeczność użytkowników, która może samodzielnie (niezależnie od innych języków) definiować pojęcie jakości w ramach swojej wersji [**102**, **166**, **186**, **189**]. Systemy oceniania mają swoje regulaminy oraz zasady przyznawania wyróżnień za jakość, ustalone w toku dyskusji i wieloletniej praktyki.

W poszczególnych wersjach Wikipedii definiowane są systemy wyróżnień dla artykułów, które uważane są za wzorowe. Takie artykuły istnieją w niemal każdej wersji językowej. W angielskiej wersji Wikipedii najlepsze artykuły nazywają się Featured Article (FA). Istnieje również drugie wyróżnienie dla artykułów, które jeszcze nie spełniają wszystkich kryteriów FA, ale zbliżają się do ich jakości – Good Article (GA). W polskiej wersji odpowiednikami FA i GA są „Artykuł na medal" i „Dobry artykuł". Nadawanie tych wyróżnień następuje w wyniku konsensusu po przeprowadzonej dyskusji, w której każdy użytkownik może się zgodzić lub nie z nadaniem artykułowi wyższej oceny jakości oraz wyjaśnić swój punkt widzenia. Zasady mogą się zmieniać

w czasie, niezależnie w każdej wersji językowej, co w konsekwencji może spowodować utratę wyróżnienia przez niektóre artykuły.

W niektórych wersjach językowych artykuł może dostać również inną ocenę jakości. Taka ocena może odzwierciedlać „dojrzałość" artykułu oraz w jakim stopniu jest on zbliżony do wzorowego. Na przykład w angielskiej wersji Wikipedii wyróżnia się 7 klas jakości artykułu (od najwyższej): FA, GA, A-class, B-class, C-class, Start, Stub. Wszystkie oceny poniżej FA i GA nadawane są bez dyskusji społeczności i uzyskania konsensusu – każdy użytkownik może wystawić ocenę samodzielnie na podstawie przyjętych zasad. W wersji polskiej, oprócz wyróżnionych artykułów, istnieją jeszcze następujące oceny: Czwórka, Start, Zalążek. Warto dodać, że w zależności od projektów tematycznych w ramach tej samej wersji językowej mogą być stosowane rożne nazwy klas o podobnych kryteriach. Np. podobną do klasy „Czwórka" w polskojęzycznej Wikipedii jest klasa o nazwie „Poprawny", a odpowiednikiem klasy startowej w poszczególnych projektach tematycznych jest klasa dostateczna.

Dodatkowym interesującym aspektem oceny jakości artykułów przez użytkowników jest fakt, iż jeden artykuł może posiadać jednocześnie różne oceny jakości, w zależności od projektu tematycznego. Na przykład w momencie przeprowadzenia badań artykuł o Poznaniu w angielskiej Wikipedii[11] jest oceniony oddzielnie z punktu widzenia pięciu projektów tematycznych: WikiProject Poland, WikiProject Middle Ages, WikiProject Cities, WikiProject Former countries / Prussia oraz WikiProject Germany. W tym przypadku oceny pomiędzy projektami są spójne: klasa C. Oceny od poszczególnych projektów zazwyczaj można sprawdzić na stronach dyskusji nad artykułem. Przykładem niespójności ocen pomiędzy projektami tematycznymi może być artykuł o Tajwanie w angielskiej Wikipedii[12], gdzie 5 projektów ocenia artykuł jako klasa B, podczas gdy 3 inne projekty przypisały klasę C. Artykuł może mieć dowolną liczbę ocen jednocześnie od różnych projektów tematycznych – np. artykuł o Google w anglojęzycznej Wikipedii[13] ma ich 10. Dodatkowo każdy projekt może wyznaczyć ocenę ważności danego artykułu.

W tabeli **7.3** pokazana jest liczebność artykułów w poszczególnych klasach jakości. Dla artykułów, które miały niespójne oceny jakości pomiędzy projektami tematycznymi Wikipedii, była wybierana najniższa.

11 Adres strony dyskusji nad artykułem o Poznaniu w angielskiej Wikipedii: https://en:wikipedia-:org/wiki/Talk:Poznan

12 Adres strony dyskusji nad artykułem o Taiwaniu w angielskiej Wikipedii: https://en:wikipedia:org/wiki/Talk:Taiwan

13 Adres strony dyskusji nad artykułem o Google w angielskiej Wikipedii: https://en:wikipedia:org/wiki/Talk:Google

TABELA 7.3. **LICZBA ARTYKUŁÓW W POSZCZEGÓLNYCH KLASACH JAKOŚCI W RÓŻNYCH WERSJACH JĘZYKOWYCH WIKIPEDII**

	WERSJA JĘZYKOWA					
	BE	**DE**	**EN**	**PL**	**RU**	**UK**
Wszystkie artykuły	155.256	2.195.346	5.674.716	**1.288.046**	1.481.270	798.850
Featured Article (FA)	66	2.531	5.344	**795**	1.119	247
Good Article (GA)	109	3998	28.082	**2.107**	3.122	649
Solidny artykuł					4.977	
A-class			2.061			
Czwórka				**167**		
Pełny artykuł					5.585	191
B-class			73.081			
Rozwinięty artykuł					18.464	1.601
C-class			202.551			
Artykuł w rozwoju					76.901	7.017
Start			1.430.317	**862**		
Zalążek (Stub)	964		2.834.390	**637**	85.906	25.187
Nieocenione	154.117	2.188.817	1.098.890	**1.283.478**	1.285.196	763.958

Źródło: opracowanie własne, stan na lipiec 2018 r.

Z zestawienia wynika, że nie ma ogólnie przyjętego standardu klasyfikacji artykułów pomiędzy różnymi wersjami językowymi Wikipedii. Niektóre języki stosują rozwiniętą skalę ocen (EN, RU), inne ograniczają się do 2-3 klas jakości (BE, DE). Poza tym, w rozwiniętych klasyfikacjach również nie ma spójności pomiędzy językami, jednak można znaleźć podobieństwa w zasadach przyznawania poszczególnych ocen (w tabeli **7.3** podobne klasy zostały pogrupowane).

Każda wersja językowa może mieć swój system klasyfikacji jakości artykułów, jednak można zauważyć, że wszystkie stosują co najmniej dwie klasy wyróżnionych artykułów – odpowiedniki FA i GA. Artykułów wyróżnionymi tymi klasami jest bardzo mało – średnio w każdej wersji językowej ich udział wynosi około 0,07%. Warto również podkreślić, że duża część artykułów nie jest w ogóle oceniona, np. w polskiej edycji udział takich artykułów stanowi ponad 99% [**102**].

Z jednej strony dane i informacje dostarczane przez społeczność mogą być kwestionowane, jak wcześniej wspomniana krytyka Wikipedii. Z drugiej strony, stworzone przez społeczność ramy oceny jakości pozwalają na poprawę jakości danych, w szczególności w części ustrukturyzowanej, tj. w infoboksach. W tym celu mogą być wykorzystane dwa mechanizmy:

- formalny proces oceny jakości artykułów Wikipedii, który dostarcza danych,
- wielojęzyczność i powiązania między artykułami w różnych językach, stwarzające szanse w zakresie porównania i weryfikacji poprawności wprowadzanych danych.

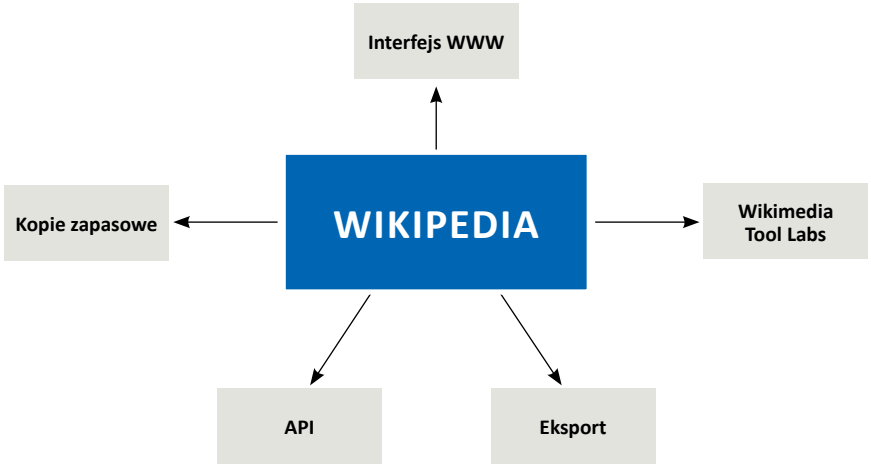
Pozyskiwanie miar jakości

W związku z wielością źródeł, z których można pozyskać miary do oceny jakości informacji, poniżej zostały przedstawione metody adekwatne do każdego ze źródeł.

Najważniejsze metody pozyskania miar do oceny jakości to: API, pobranie zrzutu stron (dump), zapytania SQL, zapytania SPARQL, przetwarzanie języka naturalnego (NLP). W przypadku NLP należy uwzględnić różne frameworki oraz wielojęzyczność, tzn. pozyskiwanie danych z wykorzystaniem charakterystycznych zasobów, np. do analizy morfologicznej.

Na rys. **7.7** pokazane są różne podejścia do pozyskiwania danych dotyczących artykułów Wikipedii. Najprostszym jest interfejs WWW. Kolejnym stosunkowo prostym narzędziem dostępu do artykułów Wikipedii jest specjalny formularz[14] pozwalający eksportować jeden lub wiele artykułów wraz z historią edycji (do 1000 edycji) w formacie XML.

RYSUNEK 7.7. **MOŻLIWOŚCI DOSTĘPU DO DANYCH ARTYKUŁÓW WIKIPEDII**



Źródło: opracowanie własne.

Wikipedia co miesiąc tworzy pełną kopię wszystkich swoich wersji językowych w postaci wikitekstu (kodu źródłowego), metadanych w różnych formatach (w tym XML) oraz surowych baz danych w postaci SQL.

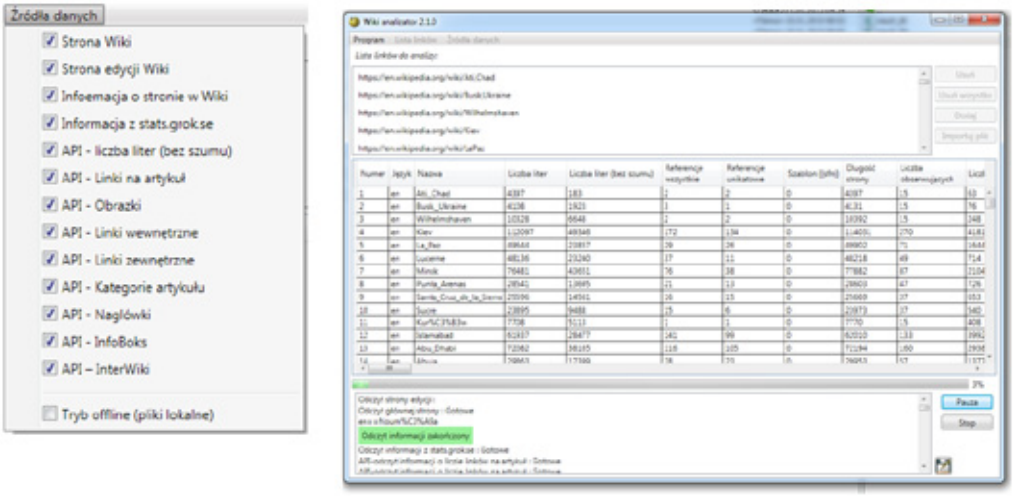
14 https://pl.wikipedia.org/wiki/Specjalna:Eksport

Wszystkie te pliki są dostępne na specjalnym serwerze[15]. Wykorzystanie kopii zapasowych do analizy dużej liczby artykułów jest bardziej czasochłonne niż inne metody. Jest to związane m.in z wielkością plików, np. dla angielskiej wersji Wikipedii jeden z plików zawierający ostatnie wersje tekstów źródłowych artykułów ma objętość ponad 70GB.

Wikipedia Tool Labs[16] (dawniej Toolserver) jest środowiskiem serwerowym oferowanym użytkownikom Wikipedii, przeznaczonym do tworzenia oprogramowania służącego do ułatwienia korzystania z tej encyklopedii. Ponad 1000 narzędzi pozwala m.in. otrzymać różnego rodzaju statystyki dotyczące artykułów. Jednak programy te nie zawsze są dostępne, mogą zawierać błędy lub nieaktualne sposoby analizy informacji.

Jedną z najbardziej atrakcyjnych metod pozyskiwania danych jest serwis API, który zapewnia wygodny dostęp m.in. do danych i metadanych artykułów Wikipedii za pomocą protokołu HTTP, za pośrednictwem adresu URL, w różnych formatach (w tym XML, JSON). W odróżnieniu od kopii zapasowych pobierane dane są aktualne w momencie zadania zapytania, a odpowiedź serwera na zapytanie jest szybka. Z możliwości API korzysta program WikiAnalizator, który może pozyskać ponad 50 różnych cech poszczególnych artykułów. Na rysunku **7.8** pokazany jest interfejs graficzny tego programu oraz źródła, z których pobiera on dane.

RYSUNEK 7.8. **INTERFEJS GRAFICZNY PROGRAMU WIKIANALIZATOR WRAZ Z WYKAZEM ŹRÓDEŁ DANYCH**



Źródło: opracowanie własne.

15 https://dumps.wikimedia.org
16 http://tools.wmflabs.org

Serwis API działa dla każdego języka Wikipedii i jest dostępny pod adresem określonym według szablonu:

https://{język}.wikipedia.org/w/api.php?action={ustawienia}

gdzie:

- {język} – dwuliterowe okerślenie wersji językowej,
- {ustawienia} – ustawienia zapytania[17].

Na przykład w celu otrzymania określonych informacji dotyczących artykułu „Polska" w języku polskim, w formacie JSON, należy wykorzystać następujące wywołania:

- kod źródłowy artykułu w notacji wiki: https://pl.wikipedia.org/w/api.php?action=parse&format=json&prop=wikitext&page=Polska
- tylko tekst artykułu: https://pl.wikipedia.org/w/api.php?format=json&action=query-&prop= extracts&explaintext&titles=Polska
- lista wszystkich nagłówków: https://pl.wikipedia.org/w/api.php?action=parse&format=json&prop=sections&page=Polska
- lista wszystkich linków wewnętrznych: https://pl.wikipedia.org/w/api.php?action=parse&format=json&prop=links&page=Polska
- lista wszystkich szablonów: https:///pl.wikipedia.org/w/api.php?action=parse&format=json&prop=templates&page=Polska
- lista wszystkich obrazów: https://pl.wikipedia.org/w/api.php?action=parse&format=json&prop=images&page=Polska

Każdy artykuł posiada dedykowaną stronę do dyskusji[18], która służy do wymiany opinii, zgłaszania uwag i rozwiązywania konfliktów związanych z treścią artykułu. Również te strony można wykorzystać do ekstrakcji miar, zadając podobne zapytania.

Wykorzystując kopie zapasowe do ekstrakcji miar jakości, należy przeanalizować następujące pliki (dla polskojęzycznej Wikipedii):

- plwiki-latest-pages-meta-current.xml.bz2: rekombinacja wszystkich stron (łącznie artykułów), tylko aktualne wersje. Ten plik służy do uzyskania większości miar artykułów.

17 Wszystkie możliwe ustawienia serwisu API można znaleźć na specjalnej stronie: https:// pl.wikipedia.org/wiki/Specjalna:ApiSandbox

18 https://pl.wikipedia.org/wiki/Pomoc:Strona_dyskusji

- plwiki-latest-pages-articles.xml.bz2: zawiera artykuły, szablony, opisy obrazów / plików i podstawowe meta-strony. Może być również wykorzystany do uzyskania większości miar artykułów (z wyłączeniem miar dotyczących stron dyskusji).
- plwiki-latest-pagelinks.sql.gz: rekordy powiązań wiki pomiędzy stronami. Używany do miar sieciowych – na przykład linki przychodzące z innych artykułów.
- plwiki-latest-categorylinks.sql.gz: rekordy opisujące przynależność stron wiki do kategorii. Może być stosowany do pomiaru liczby kategorii.
- plwiki-latest-externallinks.sql.gz: rekordy zewnętrznych adresów URL na stronach wiki. Może być użyty do pomiaru liczby linków wychodzących (zewnętrznych).
- plwiki-latest-stub-meta-history.xml.gz: zawiera tylko metadane z historii edycji stron wiki. Może być użyty do ekstrakcji liczby autorów z różnych grup(botów, anonimowych użytkowników, administratorów itp.), oraz liczby edycji różnych typów (np. drobne zmiany, komentarze do edycji).
- plwiki-latest-iwlinks.sql.gz: informacja o linkach typu interwiki. Może być użyty do wyodrębnienia liczby unikalnych linków wewnętrznych (linki do innych artykułów Wikipedii).
- plwiki-latest-templatelinks.sql.gz: rekordy użycia szablonów przez strony wiki. Używane do określenia liczby szablonów oraz do sprawdzenia, czy artykuł posiada infoboks.
- plwiki-latest-page.sql.gz: dane bazowe o stronach wiki (identyfikator, tytuł, ograniczenia itp.). Może być użyty do wyodrębnienia daty i godziny ostatniej edycji, długości strony w bajtach.
- plwiki-latest-imagelinks.sql.gz: rekordy użycia plików/obrazów na stronach wiki. Może być używany do zliczania obrazów.

Opisane powyżej pliki kopii zapasowych Wikipedii mogą dawać różne możliwości ekstrakcji miar. Na przykład liczbę obrazów można liczyć biorąc pod uwagę tag [[image: ...]] w kodzie źródłowym strony wiki [**25**, **35**, **40**, **167**, **168**, **199**, **201**, **203**]. Jednak dodatkowe obrazy, umieszczone innym sposobem (na przykład za pomocą specjalnych szablonów), nie będą wtedy brane pod uwagę. Dlatego można wykorzystać inne podejście, które wyodrębni liczbę obrazów z pliku rekordów użycia obrazów wiki [**96**, **98**, **100**, **102**, **159**, **189**].

Inny przykładem jest liczba linków wewnętrznych (interwiki) wychodzących ze strony i liczba przychodzących wewnętrznych linków do strony (z innych artykułów). W celu ekstrakcji tej miary można rozpatrywać kod źródłowy każdego artykułu, aby znaleźć linki do innych stron. Jednak w tym przypadku nie mogą być zidentyfikowane linki, które zostały wstawione przez specjalne szablony. Dlatego może być używany również plik z rekordami łącza stron wiki pomiędzy sobą.

Niektórych miar nie można wyekstrahować z plików kopii zapasowych Wikipedii. Na przykład, aby dla każdego artykułu uzyskać liczbę obserwatorów stron, konieczne jest wysyłanie

zapytania do Wikipedia API [**49**]. Miary z zasobów zewnętrznych (takich jak Facebook, Twitter, Reddit itp.) należy również ekstrahować oddzielnie z innych źródeł.

Na podstawie literatury [**9**, **24**, **25**, **40**, **41**, **77**, **98**, **99**, **100**, **107**, **109**, **167**, **186**, **198**] oraz badań własnych, w tabeli **7.4** przedstawiono listę możliwych do ekstrakcji miar. Uwzględniono różne źródła (w tym strony dyskusji nad artykułami), z których można pozyskać miary do oceny jakości informacji.

TABELA 7.4. **MIARY JAKOŚCI ARTYKUŁÓW WIKIPEDII**

SKRÓT	OPIS MIARY
A1, A2	Długość artykułu (w bajtach lub według liczby znaków)
A3, A4	Długość artykułu bez szumu (w bajtach lub według liczby znaków)
A5	Liczba obrazków
A6, A7	Liczba obrazków z kodu wiki (w tekście lub w abstrakcie)
A8	Liczba sekcji
A9	Liczba sekcji z wyjątkiem źródłowych
A10	Liczba kategorii, wpisanych do kodu
A11, A12	Liczba wszystkich szablonów (w tekście, w abstrakcie)
A13	Liczba unikatowych szablonów
A14, A15	Liczba szablonów 1-go poziomu (w tekście lub w abstrakcie)
A16, A17	Długość szablonów 1-go poziomu (w tekście lub w abstrakcie)
A18, A19	Maksymalna długość szablonu (w tekście lub w abstrakcie)
A20	Liczba szablonów o lukach jakości
A21, A22	Liczba linków wewnętrznych (w tekście lub w abstrakcie)
A23, A24	Długość tekstu bez referencji (w bajtach lub według liczby znaków)
A25	Data i godzina ostatniej zmiany artykułu
A26, A27	Długość abstraktu (w bajtach lub według liczby znaków)
A28, A29	Długość abstraktu bez szumu (w bajtach lub według liczby znaków)
A30	Minimalna długość sekcji
A31	Maksymalna długość sekcji
A32	Liczba sekcji 1-go poziomu
A33	Liczba sekcji 2-go poziomu
A34	Liczba sekcji 1-go i 2-go poziomu (razem)
A35	Liczba tabeli
A36–53	Linki przychodzące z poszczególnych przestrzeni nazw (ns0-ns3, ns6-ns15, ns100, ns101, ns828, ns829)
A54	Linki przychodzące ze wszystkich rozpatrywanych przestrzeń nazw

SKRÓT	OPIS MIARY
A55	Linki przychodzące z artykułów Wikipedii
A56	Data i godzina ostatniego przetwarzania strony
A57	Data i godzina ostatniego przetwarzania linków
A58	Numer ostatniej edycji artykułu
A59	Długość kodu źródłowego
A60	Liczba przekierowań na stronę
A61	Liczba wersji językowych
A62	Data utworzenia artykułu
A63	Liczba edycji artykułu
A64	Liczba drobnych edycji
A65-68	Liczba edycji w ciągu ostatnich 30, 90, 180, 365 dni
A69-72	Liczba drobnych edycji w ciągu ostatnich 30, 90, 180, 365 dni
A73	Liczba unikatowych autorów
A74	Liczba unikatowych autorów anonimowych
A75 78	Liczba unikatowych autorów dokonujących zmian w ciągu ostatnich 30, 90, 180, 365 dni
A79–82	Liczba unikatowych autorów anonimowych dokonujących zmian w ciągu ostatnich 30, 90, 180, 365 dni
A83	Liczba obserwujących
A84	Suma odwiedzin w ostatnim roku rok
A85	Mediana odwiedzin w ciągu ostatnich 90 dni
A86	Mediana odwiedzin w ciągu ostatnich 365 dni
A87	Mediana odwiedzin w ciągu ostatnich 365 dni bez dni z brakiem odwiedzin
A88, A89	Liczba wszystkich referencji (w tekście lub w abstrakcie)
A90, A91	Liczba unikatowych referencji (w tekście lub w abstrakcie)
A92	Gęstość referencji
A93	Długość kodu referencji
A94	Liczba referencji ze specjalnym szablonem
A95	Średnia liczba wypełnionych parametrów w szablonach referencji
A96	Liczba referencji posiadających archiwum
A97	Liczba referencji posiadających co najmniej jeden specjalny identyfikator
A98–105	Liczba referencji z określonym specjalnym identyfikatorem (DOI, ISBN, ISSN, OCLC, ARXIV, PMID, PMC, JSTOR)
A106, A107	Referencje z 50 oraz 100 najbardziej popularnych hostów w ramach wybranej wersji językowej Wikipedii
A108–110	Referencje z 50, 100 oraz 300 najbardziej popularnych hostów w ramach wybranej wersji językowej Wikipedii
A111–115	Referencje z 50 najbardziej popularnych domen określonego państwa (Białoruś, Polska, Rosja, Ukraina, USA)
A116–119	Referencje z szablonem określonego rodzaju (cytowania strony WWW, wiadomości, książki, czasopisma)

SKRÓT	OPIS MIARY
A120, A121	Długość strony dyskusji (w bajtach oraz według liczby znaków)
A122	Liczba szablonów na stronie dyskusji
A123	Liczba sekcji na stronie dyskusji
A124	Liczba linków wewnętrznych na stronie dyskusji
A125	Długość abstraktu na stronie dyskusji
A126	Minimalna długość sekcji na stronie dyskusji
A127	Maksymalna długość sekcji na stronie dyskusji
A128	Data i godzina ostatniego odświeżenia na stronie dyskusji
A129	Data i godzina ostatniego odświeżenia linków na stronie dyskusji
A130	Numer ostatniej edycji na strony dyskusji
A131	Długość kodu źródłowego strony dyskusji
A132	Suma odwiedzin w ciągu ostatnich 365 dni strony dyskusji
A133	Liczba sygnałów z serwisu Facebook do artykułu Wikipedii

Źródło: opracowanie własne.

Szczególnie warto opisać miary A_{36-53}, które są związane z liczbą linków ze stron różnych grup Wikipedii. Strony w Wikipedii są pogrupowane w tzw. przestrzenie nazw (ang. namespace lub ns). W zależności od wersji językowej liczba takich zdefiniowanych grup może się różnić. Na przykład dla polskojęzycznej wersji zdefiniowano 29 przestrzeni nazw [**145**]. Pomiędzy wersjami językowymi można znaleźć spójne oznaczenia przestrzeni nazw, które mogą okazać się przydatne do ekstrakcji miar jakości. Tabela **7.5** przedstawia oznaczenie oraz opis wybranych przestrzeni nazw stron Wikipedii. W dalszych analizach będą używane miary z wszystkich przedstawionych przestrzeni nazw, z wyjątkiem ns4 oraz ns5, które zawierają informację ogólne na temat Wikipedii. Takie strony często zawierają linki do stron, które mają przypisane najwyższe oceny jakości – FA oraz GA. Jest to związane m.in. z tym, że w przestrzeni nazw ns4 umieszczane są listy najlepszych artykułów. Oznacza to, że artykuły z treścią podobnej jakości nie będą mieć linków z tych stron, co może prowadzić do zmniejszenia precyzji modeli jakości podczas ewaluacji nieocenionych artykułów.

TABELA 7.5. **SKRÓTY ORAZ OPISY WYBRANYCH PRZESTRZENI NAZW**

OZNACZENIE	OPIS PRZESTRZENI NAZW
ns0	Artykuły encyklopedyczne - główna przestrzeń nazw
ns1	Dyskusja artykułu
ns2	Wikipedyst(k)a- strony użytkowników Wikipedii, na podstronach - pomocnicze strony do własnego użytku, przyborniki, brudnopisy
ns3	Dyskusja wikipedysty/-ki
ns4	Wikipedia- informacje ogólne na temat Wikipedii, zasady oraz różnego rodzaju strony współpracy i dokumentacji
ns5	Dyskusja Wikipedii
ns6	Plik- strony opisujące załadowane pliki multimedialne (obraz, dźwięk, film)
ns7	Dyskusja pliku
ns8	MediaWiki - komunikaty interfejsu.
ns9	Dyskusja MediaWiki
ns10	Szablon - strony przeznaczone do definiowania oraz opisywania szablonów
ns11	Dyskusja szablonu
ns100	Portal- strony prezentujące czytelnikom w atrakcyjny sposób tematyczny wycinek zawartości Wikipedii
ns101	Dyskusja portalu
ns828	Moduł - strony zawierające kod w języku Lua, współpracujące ze skomplikowanymi szablonami
ns829	Dyskusja modułu

Źródło: opracowanie własne na podstawie [145].

Modele do automatycznego określania jakości

Niniejsza sekcja przedstawia przykładowy sposób oceny jakości danych o nietypowej charakterystyce (tekstowych, społecznościowych), z wykorzystaniem miar powszechnie stosowanych w ML oraz definicjami miar unikalnych dla domeny.

Na podstawie ocenionych artykułów Wikipedii można zbudować predykcyjny model jakości. W tym przypadku, na podstawie znajomości wartości rożnych miar, model musi nauczyć się określać przynależność nieocenionych artykułów do określonych klas jakości. Wówczas w takim modelu zmienna zależna będzie nominalna, czyli przyjmowała pewien określony skończony zbiór wartości (niekoniecznie liczbowych).

Artykuły można również dobierać w grupy klas. Wtedy liczba kategorii będzie redukowana. Na potrzeby budowy modeli będą używane dwie grupy klas jakości: „Kompletne” oraz

„Niekompletne". W takim modelu zmienna zależna będzie dychotomiczna, tzn. będzie mogła przyjmować tylko jedną z dwóch wartości.

Model oceny jakości może być zbudowany na podstawie zmiennej zależnej, która odpowiada klasom jakości określonych przez użytkowników Wikipedii. W zależności od wersji językowej Wikipedii liczba takich klas w zbiorze danych może być różna. Np. dla angielskiej Wikipedii jest 6 wartości zmiennej zależnej: FA, GA, B-class, C-class, Start, Stub. Artykuły z klasy A zazwyczaj posiadają dodatkowo wyższą ocenę FA bądź GA, dlatego ta klasa nie będzie rozpatrywana oddzielnie, podobnie jak to było robione w innych badaniach [**43**, **67**, **160**].

W celu zbudowania zbilansowanej próby, liczba artykułów z poszczególnych klas była dobrana biorąc pod uwagę liczebność najmniejszej klasy jakości – FA, która zawierała 5.344 artykuły. W wyniku losowego doboru 5.000 artykułów z każdej klasy otrzymano zbiór 30.000 artykułów.

Niektóre algorytmy wymagają konwersji wartości kategorialnej zmiennej zależnej na liczby (np. algorytmy klasyfikacji w bibliotece scikit-learn19 w języku programowania Python). W takim przypadku ważne jest, aby zachować kolejność klas zgodnie z malejąca lub rosnącą jakością [**42**].

Modele jakości dwuklasowe

Do zbudowania modelu została zastosowana dychotomiczna zmienna objaśniana [**99**, **104**, **170**, **186**, **200**], gdzie jakość jest modelowana jako prawdopodobieństwo przynależności do jednej z dwóch klas:

- Kompletne artykuły: klasy FA i GA,
- Niekompletne artykuły: wszystkie inne – rozwijające się (które należy dopracować) oraz nieocenione artykuły.

Można również zastosować podejście podziału na artykuły klasy FA (jako wzorowych) oraz innych losowo dobranych artykułów [**170**].

Dobór próby uczącej również musi się odbywać z uwzględnieniem liczby artykułów w najmniejszej klasie. W związku z tym wybrano losowo 5.000 artykułów z klasy FA oraz 5.000 artykułów z klasy GA do umieszczenia w kategorii „Kompletne", co dało łącznie 10.000 artykułów. Taką samą liczbę artykułów wylosowano do kategorii „Niekompletne", przyczym w celu zachowania równoważności tej grupy, z pozostałych 4 klas jakości (B,C, Start, Stub) wylosowano

19 https://scikit-learn.org

odpowiednio 2.500 artykułów. Cała próba ucząca wykorzystana do budowania modeli jakości przy użyciu dychotomicznej zmiennej zależnej wyniosła 20.000 artykułów.

Do oceny jakości klasyfikatorów wykorzystywane mogą być różne narzędzia. Jednym z podstawowych jest macierz błędu, która może być stosowana do pokazania rozbieżności klasyfikacji – klasy, do których należą artykuły (klasy rzeczywiste) oraz klasy, które były określone przez model (wynik). Innymi słowy ta macierz wskazuje, ile z oryginalnie oznaczonych jako „Kompletne" („Niekompletne") zostanie omyłkowo zaklasyfikowana jako „Niekompletne" („Kompletne"). Tabela **7.6** pokazuje macierz błędów tego modelu oceny jakości przy wykorzystaniu algorytmu lasu losowego (Random Forest).

TABELA 7.6. **MACIERZ BŁĘDÓW W MODELU PREDYKCJI JAKOŚCI W ANGIELSKIEJ WIKIPEDII DLA DWÓCH KLAS JAKOŚCI Z WYKORZYSTANIEM ALGORYTMU RANDOMFOREST**

		Wynik modelu	
		Kompletne	Niekompletne
Klasa rzeczywista	Kompletne	9615	385
	Niekompletne	804	9196

Źródło: obliczenia własne przy użyciu pakietu statystycznego WEKA.

Na podstawie macierzy błędów można obliczyć różne wskaźniki jakości modelu. Niektóre z tych wskaźników zostały opisane w tabeli **7.7.**

Tabela **7.8** pokazuje szczegółową informację na temat wskaźników jakości w modelu predykcji jakości w angielskiej Wikipedii przy użyciu dychotomicznej zmiennej zależnej z wykorzystaniem algorytmu RandomForest.

Modele jakości wieloklasowe

Wieloklasowe modele jakości są budowane przy zachowaniu oryginalnych klas jakości. W angielskiej Wikipedii takich klas jest 6: FA, GA, B, C, Start, Stub. Dla każdej z tych klas zostało losowo dobrane po 5.000 artykułów. Razem zbiór danych liczył 30.000 artykułów. Dla każdego z wybranych artykułów zostały wyekstrahowane ponad 100 różnych miar jakości, opisanych w rozdziale „Miary oraz wymiary jakości artykułów Wikipedii".

TABELA 7.7. **WSKAŹNIKI JAKOŚCI MODELU KLASYFIKACYJNEGO**

WSKAŹNIK	OPIS
TP (True Positive)	Prawdziwie pozytywna. W rozpatrywanym przykładzie jest to wartość 9615, która pokazuje ile artykułów oryginalnie oznaczonych jako „Kompletne" zostały zaklasyfikowany jako„Kompletne" przez model. TP Rate oznacza stosunek artykułów oznaczonych przez model jako „Kompletne" do liczby wszystkich artykułów, które rzeczywiście do tej klasy należą.
TN (True Negative)	Prawdziwie negatywna. W rozpatrywanym przykładzie jest to wartość 9196, która pokazuje ile artykułów oryginalnie oznaczonych jako „Kompletne" zostały zaklasyfikowany jako „Niekompletne".
FP (False Positive)	Fałszywie pozytywna czy błąd pierwszego rodzaju. W rozpatrywanym przykładzie jest to wartość 804, która pokazuje ile artykułów oryginalnie oznaczonych jako „Niekompletne" zostały omyłkowo zaklasyfikowany jako „Kompletne". FP Rate oznacza stosunek artykułów omyłkowo oznaczonych przez model jako „Kompletne" do liczby wszystkich artykułów, które należą do klasy „Niekompletne".
FN (False Negative)	Fałszywie negatywna czy błąd drugiego rodzaju. W rozpatrywanym przykładzie jest to wartość 385, która pokazuje ile artykułów oryginalnie oznaczonych jako „Kompletne" zostały zaklasyfikowany omyłkowo jako „Niekompletne" przez model.
Precision	Precyzja modelu, liczona na podstawie wzoru: $Precision = TP/(TP + FP)$
Recall	Czułość modelu, liczona na podstawie wzoru: $Recall = TP/(TP + FN)$
F-measure	Miara liczona na podstawie wzoru: $\frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$
MCC	współczynnik korelacji Matthews liczony na podstawie wzoru: $MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
ROC (Receiver Operating Characteristics)	Prawdopodobieństwo, że badany model predykcyjny oceni wyżej losowy element klasy pozytywnej („Kompletne") od losowego elementu klasy negatywnej („Niekompletne"). ROC- to funkcja punktu odcięcia, przedstawia zmienność TP Rate w zależności od FP Rate.
PRC (Precision-Recall Curve)	Pokazuje zależność między precyzją (Precision) a czułością (Recall) dla każdego możliwego odcięcia na wykresie, gdzie na osi OX pokazana precyzja oraz na osi OY- czułość modelu. W odróżnieniu od ROC, ta wartość może być bardziej przydatna, jeżeli badane jest zachowanie klasyfikatora tylko w ramach określonej klasy [151].

Źródło: opracowanie własne.

Podobnie jak w przypadku klasyfikacji binarnej, dla większej niż 2 liczby klas można zbudować macierz błędów, która pokazuje rozbieżności pomiędzy rzeczywistymi klasami oraz klasami, określonymi przez model. Tabela **7.9** przedstawia taką macierz dla 6 klas jakości w angielskiej Wikipedii przy użyciu algorytmu lasu losowego. Warto zwrócić uwagę, że największe rozbieżności pomiędzy liczbami artykułów rzeczywistych oraz oznaczonych przez model, występują w sąsiadujących pod względem jakości klasach. Może to oznaczać brak sztywnych granic pomiędzy kryteriami jakości w bliskich klasach. Najniższa predykcja modelu wykazana przy wyznaczeniu jakości artykułów z pośrednich klas: B oraz C.

TABELA 7.8. **WSKAŹNIKI JAKOŚCI W MODELU PREDYKCJI JAKOŚCI W ANGIELSKIEJ WIKIPEDII DLA DWÓCH KLAS JAKOŚCI Z WYKORZYSTANIEM ALGORYTMU RANDOMFOREST**

KLASA	TP RATE	FP RATE	PRECISION	RECALL	F-MEASURE	MCC	ROC AREA	PRC AREA
Kompletne	0,962	0,08	0,923	0,962	0,942	0,882	0,985	0,982
Niekompletne	0,92	0,039	0,96	0,92	0,939	0,882	0,985	0,986
Średnia ważona	0,941	0,059	0,941	0,941	0,941	0,882	0,985	0,984
Śr. ważona	0,648	0,07	0,642	0,648	0,642	0,574	0,913	0,688

Źródło: obliczenia własne przy użyciu pakietu statystycznego WEKA.

TABELA 7.9. **MACIERZ BŁĘDÓW W MODELU PREDYKCJI JAKOŚCI W ANGIELSKIEJ WIKIPEDII DLA WIELU KLAS JAKOŚCI Z WYKORZYSTANIEM ALGORYTMU RANDOMFOREST**

		WYNIK MODELU					
		FA	GA	B	C	Start	Stub
KLASA RZECZYWISTA	FA	**4648**	284	60	8	0	0
	GA	970	**3657**	226	119	27	1
	B	295	657	**2141**	1244	610	53
	C	71	348	1206	***2185***	1115	75
	Start	12	77	290	974	**2917**	730
	Stub	0	1	20	107	991	**3881**

Źródło: obliczenia własne przy użyciu pakietu statystycznego WEKA.

W tym przypadku macierz błędów również może być stosowana do obliczenia miar jakości modelu. Do obliczenia wskaźników (np. ROC), które zazwyczaj są stosowane do oceny algorytmów klasyfikacji binarnej, wykorzystana jest średnia ważona, która uwzględnia obliczenia tych wskaźników z punktu widzenia poszczególnych klas. Np. przy obliczeniu ROC dla klasy FA, wszystkie artykuły są dzielone na dwie grupy: FA oraz inne klasy (jako jedna wspólna).

Tabela **7.10** przedstawia wskaźniki jakości w modelu predykcji jakości w angielskiej Wikipedii przy użyciu nominalnej zmiennej zależnej z wykorzystaniem algorytmu RandomForest.

TABELA 7.10. **WSKAŹNIKI JAKOŚCI W MODELU PREDYKCJI JAKOŚCI W ANGIELSKIEJ WIKIPEDII DLA WIELU KLAS JAKOŚCI Z WYKORZYSTANIEM ALGORYTMU RANDOMFOREST**

KLASA	TP RATE	FP RATE	PRECISION	RECALL	F-MEASURE	MCC	ROC AREA	PRC AREA
FA	0,93	0,054	0,775	0,93	0,845	0,816	0,984	0,916
GA	0,731	0,055	0,728	0,731	0,73	0,675	0,95	0,818
B	0,428	0,072	0,543	0,428	0,479	0,393	0,852	0,506
C	0,437	0,098	0,471	0,437	0,453	0,349	0,845	0,459
Start	0,583	0,11	0,515	0,583	0,547	0,451	0,876	0,547
Stub	0,776	0,034	0,819	0,776	0,797	0,758	0,969	0,885
Śr. ważona	**0,648**	**0,07**	**0,642**	**0,648**	**0,642**	**0,574**	**0,913**	**0,688**

Źródło: obliczenia własne przy użyciu pakietu statystycznego WEKA.

W podobny sposób można zbudować modele dla innych wersji językowych, jednak jakość tych modeli w dużej mierze zależy od wielkości próby ocenionych artykułów. Zgodnie z tabelą 7.3, wystarczającą liczbę takich artykułów posiada również rosyjskojęzyczna Wikipedia, która zawiera stosunkowo dużo różnych klas jakości. Dodatkowo nie wszystkie zdefiniowane klasy jakości w rosyjskojęzycznej Wikipedii pokrywają się z ocenami w angielskiej Wikipedii.

Określanie ważności miar jakości

W związku z tym, że algorytm lasu losowego tworzy wiele drzew klasyfikacyjnych, dla każdego z tych drzew konkretna miara może mieć różny wskaźnik Giniego. Można obliczyć wartość średnią tego wskaźnika dla poszczególnych zmiennych, tym samy określając ich ważność w danym modelu.

Innym sposobem obliczania ważności miary jest uwzględnienie liczby węzłów drzew klasyfikacyjnych, które tę miarę używają. Każde drzewo w algorytmie lasu losowego budowane jest na innym, losowo wybranym podzbiorze zbioru danych uczących, a wybierany jest najlepszy podział z losowego podzbioru predyktorów (miar).

Dla obliczenia ważności *W* miary jakości *m* można wykorzystać wskaźnik będący iloczynem średniego wskaźnika Gini'ego dla tej miary oraz liczby węzłów drzew klasyfikacyjnych używających tę miarę – *L*(*m*):

$$W(m) = \text{Gini}(m) \cdot L(m) \qquad (7.1)$$

W celu wygodniejszego porównywania ważności miar w różnych modelach wartości

W(*m*) są normalizowane w taki sposób, że maksymalna wartość tego wskaźnika w ramach rozpatrywanego modelu może wynosić 100 (minimalna 0).

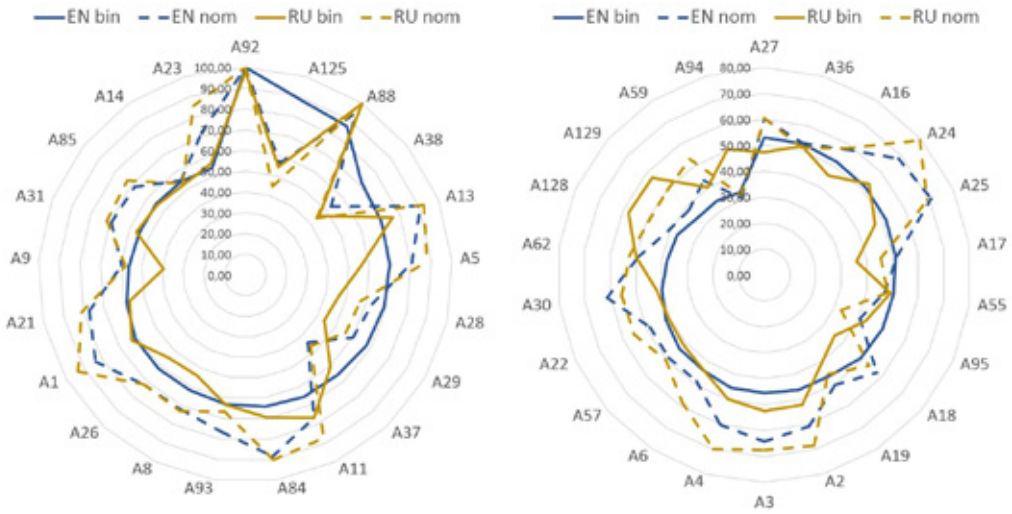
Tabela 7.11 przedstawia miary jakości (zmienne niezależne), które w skali od 0 do 100 uzyskały ważność na poziomie co najmniej 50 w co najmniej jednym z 4 modeli jakości:

- **EN bin** – dwuklasowy model jakości dla angielskiej Wikipedii.
- **EN nom** – wieloklasowy model jakości dla angielskiej Wikipedii.
- **RU bin** – dwuklaoswy model jakości z dychotomiczną (binarną) zmienną zależną w rosyjskiej Wikipedii.
- **RU nom** – wieloklasowy model jakości dla rosyjskiej Wikipedii.

Wyniki analizy ważności miar pokazują, że dla wszystkich modeli bardzo istotnym są miary związane z wiarygodnością, a w szczególności miara dotycząca gęstości referencji oraz absolutna liczba wszystkich referencji w artykule. Wskazanie źródła informacji jest więc przesłanką do wysokiej oceny jej jakości.

Wykres radarowy na rys. **7.9** pokazuje różnice pomiędzy czterema rozpatrywanymi modelami jakości.

RYSUNEK 7.9. **WAŻNOŚĆ WYBRANYCH MIAR W MODELACH PREDYKCJI JAKOŚCI W ANGIELSKIEJ (EN) LUB ROSYJSKIEJ (RU) WIKIPEDII PRZY UŻYCIU DYCHOTOMICZNEJ (BIN) LUB NOMINALNEJ (NOM) ZMIENNEJ ZALEŻNEJ Z WYKORZYSTANIEM ALGORYTMU RANDOMFOREST**



Źródło: opracowanie własne.

Syntetyczna miara jakości

Zbudowane modele mogą pozwolić na ocenę oraz porównywanie jakości w ramach jednej wybranej wersji językowej. Dla każdego języka musi zostać zbudowany odrębny model, który będzie uwzględniał różne klasy jakości. Problem może się pojawi w przypadku porównania jakości artykułów pomiędzy różnymi wersjami językowymi. Uproszczenie różnych ocen tylko do dwóch klas (zmienna dychotomiczna) jest tylko częściowym rozwiązaniem problemu. Tracimy bowiem wiele informacji z rozróżniania

TABELA 7.11. **WYBRANE WAŻNIEJSZE MIARY W MODELACH PREDYKCJI JAKOŚCI W ANGIELSKIEJ (EN) LUB ROSYJSKIEJ (RU) WIKIPEDII PRZY UŻYCIU DYCHOTOMICZNEJ (BIN) LUB NOMINALNEJ (NOM) ZMIENNEJ ZALEŻNEJ Z WYKORZYSTANIEM ALGORYTMU RANDOMFOREST**

MIARY JAKOŚCI		MODEL JAKOŚCI			
SKRÓT	OPIS	EN BIN	EN NOM	RU BIN	RU NOM
A92	**Gęstość referencji**	100,00	100,00	97,83	100,00
A125	**Długość abstraktu na stronie dyskusji**	88,87	56,44	54,56	44,80
A88	**Liczba wszystkich referencji w tekście**	86,73	94,55	100,00	99,97
A38	**Linki przychodzące z przestrzeni nazw ns2**	71,83	52,79	45,44	44,20
A13	**Liczba unikatowych szablonów**	70,49	90,67	76,25	92,63
A5	**Liczba obrazków**	69,81	80,54	55,97	87,91
A28	**Długość abstraktu bez szumu w bajtach**	68,78	62,89	46,31	57,08
A37	**Linki przychodzące z przestrzeni nazw ns1**	65,54	44,47	60,30	46,53
A11	**Liczba wszystkich szablonów w tekście**	65,31	76,58	76,57	87,33
A84	**Suma odwiedzin za ostatni rok**	64,05	88,47	69,41	90,21
A93	**Długość kodu referencji**	63,02	76,51	63,45	66,54
A8	**Liczba sekcji**	61,76	72,07	53,69	72,75
A1	**Długość artykułu w bajtach**	61,19	83,54	63,12	93,09

Źródło: obliczenia własne przy użyciu WEKA.

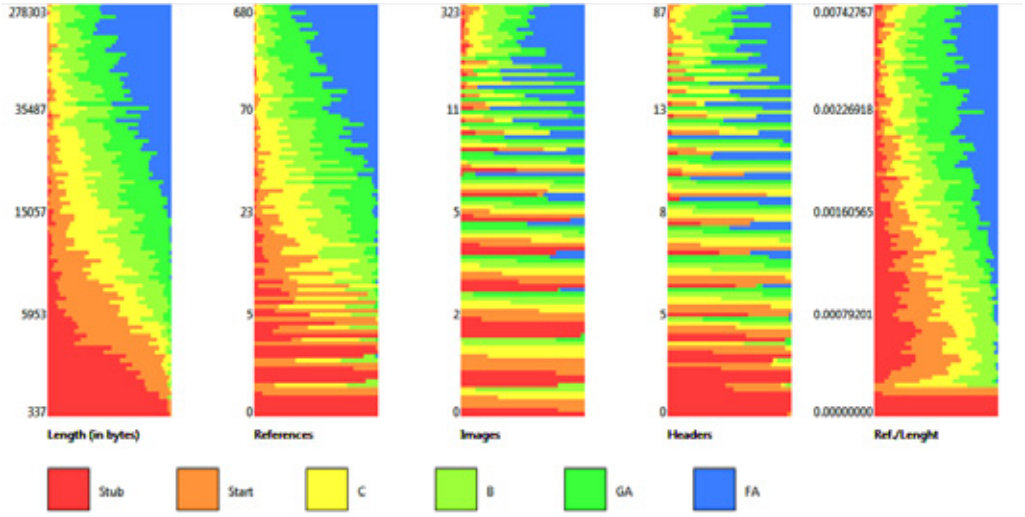
poszczególnych klas – większość klas byłaby przypisana do jednej klasy, określonej wcześniej jako „Niekompletna". Rozwiązaniem problemu mogłoby być charakteryzowanie jakości artykułu na skali ciągłej.

W celu dokładniejszej reprezentacji jakości artykułów oraz jej porównania pomiędzy wersjami językowymi można zastosować miarę syntetyczną. Jest to wartość będąca wynikiem połączenia innych oszacowań poszczególnych miar (kryteriów czy cech). Na przykład taki wskaźnik może reprezentować sumę punktów ocen ekspertów. Miara syntetyczna jest często stosowana w gospodarce i może opisać m.in. obiekt gospodarczy czy system gospodarczy jako całość [**2**, **144**, **161**].

Miara syntetyczna może pozwolić na ocenę artykułów Wikipedii w postaci zmiennej ciągłej [**102**]. Przed zbudowaniem miary syntetycznej przeprowadzono wstępną analizę wybranych miar jakości. Z 6 klas jakości zostało losowo wybrano 1000 artykułów (z każdej klasy) i dla każdego z nich zostało wyekstrahowane pięć miar: długość artykułu, liczba referencji, liczba obrazów, liczba sekcji, gęstość referencji. Wyniki prac [**100**, **186**, **189**] oraz własne eksperymenty pokazały, że te miary znajdują się wśród najważniejszych wykorzystywanych w modelach oceny jakości.

Rysunek 7.10 pokazuje rozkład wartości miar wraz udziałem artykułów z poszczególnych klas jakości. Można zauważyć, że im większa wartość miary (odłożona na osi pionowej), tym większy jest udział artykułów o wysokiej jakości. Przykładowo, jeżeli bierzemy pod uwagę miarę długość artykułów, można się spodziewać, że im dłuższy jest artykuł, tym większe prawdopodobieństwo posiadania przez niego wyższej klasy jakości.

RYSUNEK 7.10. **ROZKŁAD WYBRANYCH MIAR W ARTYKUŁACH KAŻDEJ KLASY JAKOŚCI W ANGIELSKIEJ WIKIPEDII (FA – NAJWYŻSZA KLASA, STUB – NAJNIŻSZA)**



Źródło: [102]

Pokazane zależności prowadzą do ogólnego wniosku, że większa wartość skumulowanej miary syntetycznej, która łączy najważniejsze miary, zostanie przypisana dla bardziej rozwiniętych artykułów (czyli tych artykułów, które posiadają wyższą ocenę jakości).

Dodatkowy aspekt, który był wzięty pod uwagę to różnica w kryteriach oceny artykułów w każdej wersji językowej Wikipedii. Na przykład w konkretnej wersji językowej użytkownicy mogą przykładać większą wagę do liczby referencji niż do liczby obrazów przy decydowaniu o nadaniu wysokiej oceny jakości. Poza tym wystarczająca liczba referencji do nadania określonej oceny jakości też jest określana różnie w zależności od wersji językowej. Dlatego wybrane miary jakości najlepszych (wzorcowych) artykułów należy zbadać oddzielnie dla każdej wersji językowej.

TABELA 7.12. **MEDIANY WARTOŚCI MIAR W NAJWYŻSZEJ KLASIE JAKOŚCI W RÓŻNYCH WERSJACH JĘZYKOWYCH WIKIPEDII**

JĘZYK	DŁUGOŚĆ	REFERENCJE	OBRAZY	SEKCJE	REF./DŁUGOŚĆ
BE	198 365	210	36	27	0,001106
EN	49 038	115	13	14	0,002364
PL	**59 672**	**96**	**17**	**17**	**0,001663**
RU	139 415	163	24	22	0,001169
UK	82 371,5	40,5	24,5	21	0,000491

Źródło: obliczenia własne.

Najwyższa klasa jakości (FA – w angielskiej, ANM – w polskiej) jest obecna w każdej z rozpatrywanych wersji językowych. Artykuły mogą otrzymać taką ocenę z czasem, kiedy zawartość będzie odpowiadała określonym kryteriom, np. będzie zawierała wystarczającą liczbę referencji, obrazów, sekcji etc. Można powiedzieć, że artykuł dąży do określonego progu, w którym może dostać najwyższą ocenę. Taki próg może mieć każda z rozpatrywanych miar. W celu obliczenia tych progów zostały wyekstrahowane miary wszystkich artykułów z najwyższej klasy z każdej z rozpatrywanych wersji językowych. Następnie została obliczona mediana dla każdej miary w każdym języku. Wyniki pokazane są w tabeli **7.12**.

Obliczone mediany miar będą stanowiły podstawę do normalizacji tych miar przed obliczeniem wskaźnika syntetycznego. Innymi słowy, ta mediana stanowi próg, który pokazuje stopień rozwoju artykułu według określonej miary. Przy tym, jeżeli wartość wybranej miary jest wyższa niż odpowiednia mediana (czy odpowiedni próg), to wartość znormalizowana będzie równa 1. Znormalizowana miara *i* jest zatem obliczana według wzoru:

$$\begin{cases} \frac{m_i}{p_i}, & m_i < p_i \\ 1, & m_i \geq p_i \end{cases}, \tag{7.2}$$

gdzie m_i to absolutna wartość miary *i* oraz p_i to mediana miary *i* w najwyższej klasie jakości danej wersji językowej

Zakładamy, że wszystkie miary mogą mieć podobny wpływ na jakość, a w związku z tym muszą mieć równy wpływ na wartość wskaźnika syntetycznego. Najpierw należy obliczyć średnią znormalizowanych miar (SZM) według następującego wzoru:

$$SZM = \frac{1}{c}\sum_{i=1}^{c} nm_i, \tag{7.3}$$

gdzie nm_i to znormalizowana miara *i* oraz *c* pokazuje liczbę miar.

Następnie bierzemy pod uwagę liczbę szablonów wad jakości (*SWJ*) w rozpatrywanym artykule (jeśli istniały). Biorąc pod uwagę powyższe założenia, wskaźnik syntetyczny, określający jakość artykułu, będzie liczony na podstawie wzoru:

$$\boldsymbol{Jakosc = SZM - SZM \cdot 0{,}05 \cdot SWJ,} \tag{7.4}$$

gdzie *SZM* średnią znormalizowanych miar oraz *SWJ* pokazuje liczbę szablonów wad jakości.

W tym przypadku wartość wagi 0,05 dla liczby szablonów o lukach jakości dobrana została na podstawie badań [**9**] oraz własnych obserwacji, które pokazały, że artykuły które posiadają co najmniej jeden taki szablon, nie mogą mieć wartości miary syntetycznej wyższej niż mediana liczby punktów najlepszych artykułów Wikipedii w danej wersji językowej obliczonej według wzoru 7.3. Na przykład, jeżeli artykuł posiada wszystkie wartości miar *i* wyższe niż odpowiednie mediany wartości miar p_i, to wartość *SZM* będzie maksymalna (czyli 1). W przypadku, jeżeli artykuł posiada dwa szablony wskazujący na luki w jakości, to zgodnie ze wzorem 7.4 ta wartość będzie obniżona o 0,1, co w wyniku da wartość 0,9.

Zmienna ciągła pozwala na utworzenie dowolnej liczby klas (ocen) jakości. Na przykład, jeżeli wartości wskaźnika syntetycznego zaokrąglić do części dziesiątych, możemy otrzymać 11 ocen z zakresu: 0, 0,1, 0,2, ... , 0,9, 1. Z wykorzystaniem takiej skali ocen zostały sklasyfikowane artykuły o polskich miastach w czterech wersjach językowych Wikipedii, a wyniki przedstawione w tabeli 7.13. Należy w szczególności zwrócić uwagę, że najwyżej oceniane artykuły o polskich miastach napisane są w języku polskim.

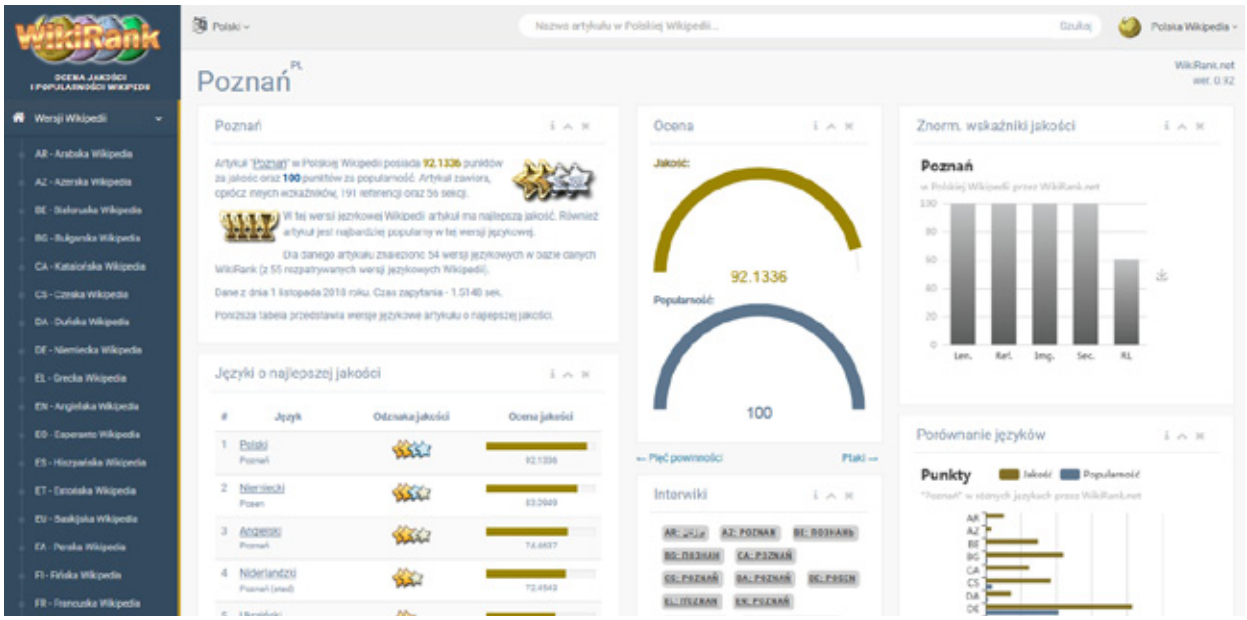
TABELA 7.13 **ZAOKRĄGLONE WARTOŚCI WSKAŹNIKA SYNTETYCZNEGO DLA ARTYKUŁÓW O POLSKICH MIASTACH W 4 WERSJACH JĘZYKOWYCH WIKIPEDII**

	ZAOKRĄGLONA WARTOŚĆ WSKAŹNIKA SYNTETYCZNEGO										
WERSJA JĘZYKOWA	**0,0**	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**	**1,0**
EN – angielska	0	142	249	233	112	100	25	19	12	5	4
PL – polska	**0**	**0**	**0**	**2**	**59**	**272**	**307**	**132**	**69**	**35**	**25**
RU – rosyjska	0	594	207	67	15	13	3	1	1	0	0
UK – ukraińska	17	414	318	81	42	12	9	5	1	2	0

Źródło: obliczenia własne.

Wyniki określenia jakości ponad 37 mln artykułów Wikipedii w różnych wersjach językowych przy użyciu miary syntetycznej dostępne są w ramach projektu WikiRank[20]. Przykład oceny jakości artykułu o Poznaniu w ramach tego projektu przedstawiony jest na rys. **7.11**.

RYSUNEK 7.11. **MIERZENIE JAKOŚCI ARTYKUŁU O POZNANIU W POLSKOJĘZYCZNEJ WIKIPEDII W RAMACH PROJEKTU WIKIRANK**

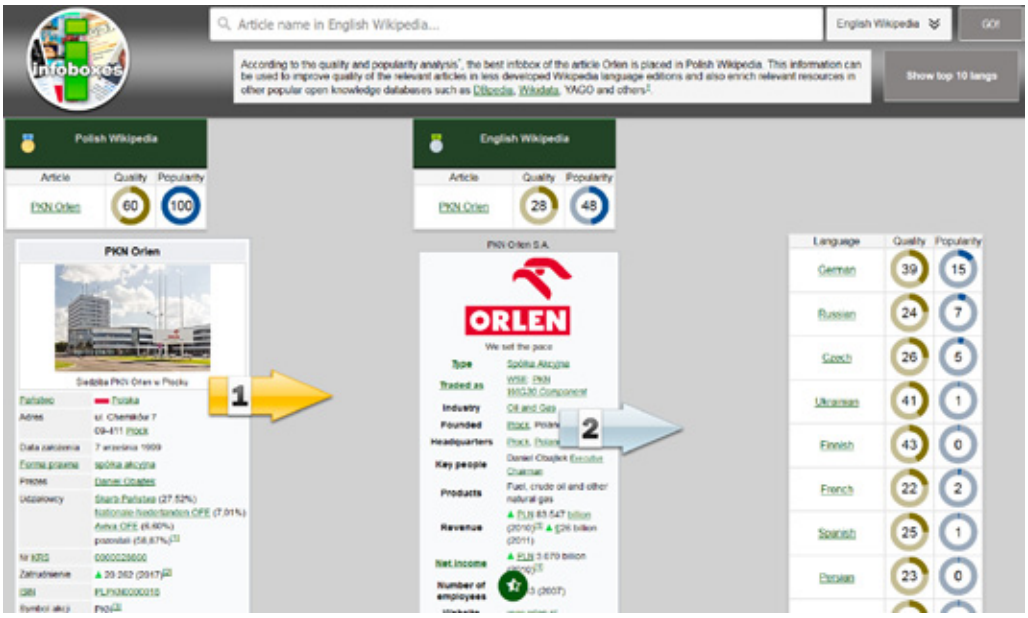


Źródło: https://pl.wikirank.net/pl/Poznań

20 https://pl.wikirank.net

Połączenie wskaźników syntetycznych jakości oraz popularności serwisu WikiRank daje możliwość porównania również części strukturalnych artykułów – tzw. infoboksów. Najlepsze dane mogą być wykorzystane m.in. do wzbogacenia mniej rozwiniętych wersji językowych Wikipedii. Na przykład, serwis Infoboxes.net[21] pokazuje, które wersje językowe należy brać pod uwagę w pierwszej kolejności, analizując dane na określony temat (rys. **7.12**).

RYSUNEK 7.12. **MIERZENIE JAKOŚCI INFOBOKSU OPISUJĄCEGO PKN ORLEN W RÓŻNYCH WERSJACH JĘZYKOWYCH WIKIPEDII**



Źródło: http://infoboxes.net/en/PKN_Orlen

7.4.2. OpenRefine

OpenRefine[22] (dawniej Google Refine) to wszechstronne narzędzie do pracy z różnymi danymi. Narzędzie pozwala m.in. na:

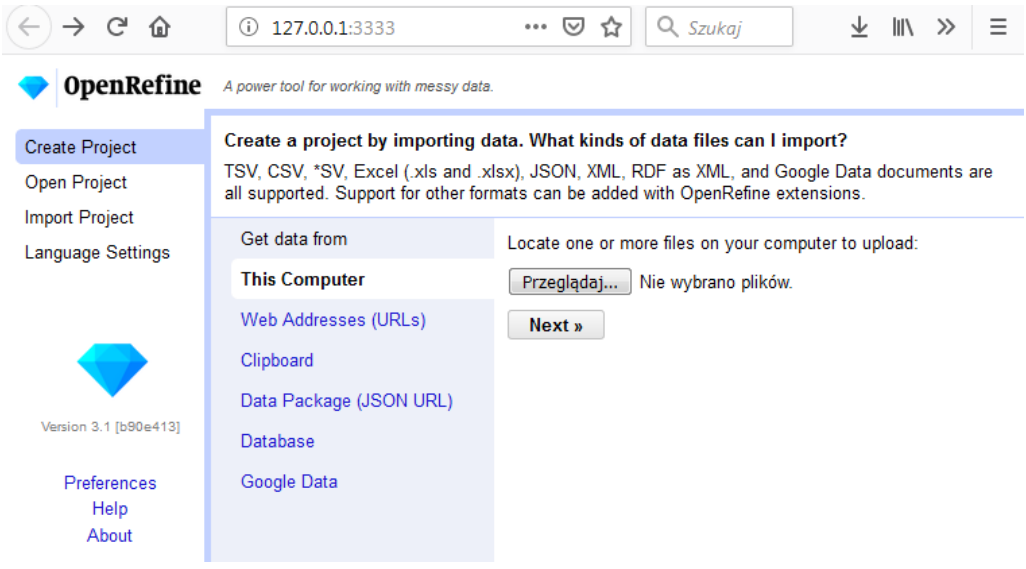
- czyszczenie danych: można uzupełniać brak danych lub poprawiać błędne wartości, a same dane można wcześniej zgrupować dla przyspieszenia pracy;
- transformację danych: konwersja wartości do innych formatów, normalizacja;

21 http://infoboxes.net
22 http://openrefine.org/

- parsowanie danych z witryn internetowych: pobieranie danych z określonych adresów URL oraz wyciągnięcie interesujących fragmentów strony WWW (parser HTML);
- dodawanie danych do zbioru danych przez pobieranie z serwisów internetowych: odbywa się to poprzez łączenie z usługami API, które zwracają pliki JSON, np. można przypisać współrzędne geograficzne adresom;
- powiązanie z dostępnymi online bazami wiedzy: zamiana danych tekstowych na odpowiednie linki do stron powiązanych, np. do treści z Wikidanych.

W celu rozpoczęcia pracy z OpenRefine należy pobrać aplikację dla wybranego systemu operacyjnego (Windows, Mac, Linux) ze strony: http://openrefine.org/download.html. W przypadku wersji dla Windows należy rozpakować pobrane archiwum i uruchomić plik „refine.bat". W nowym oknie przeglądarki, pod adresem http://127.0.0.1:3333 pojawi się główna strona aplikacji (rys. **7.13**).

RYSUNEK 7.13. **STRONA GŁÓWNA APLIKACJI OPENREFINE**



Źródło: zrzut ekranu.

Przykład 1. Czyszczenie danych

Rozpatrzmy poniższą tabelę z problematycznymi wierszami. Niektóre wiersze zawierają tylko jedną unikalną cechę, występują braki danych, a identyfikatory się powtarzają.

CECHA2	IDENTYFIKATOR	CECHA	ODNOŚNIK
	A		URL3
określona	A	określona	URL1
	A		URL2
określona	B	określona	URL2
	B		URL1
	B		URL3

Rezultat, który chcemy osiągnąć:

- połączyć w jednej komórce odnośniki dla każdego unikatowego identyfikatora, rozdzielone przecinkiem,
- usunąć duplikaty wierszy.

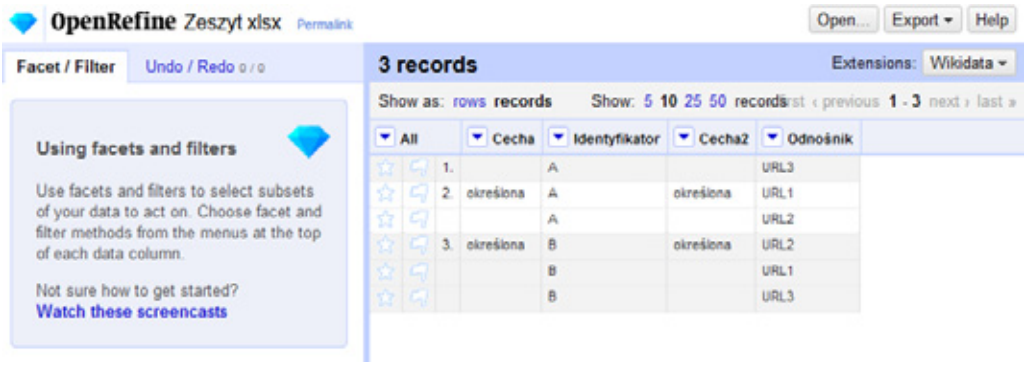
Tabela została zapisana w formacie Excel (.xlsx) i załadowana do systemu wykorzystując okno startowe systemu. Po kliknięciu „Next" otworzy się okno z otworzoną tabelą wraz z dodatkowymi narzędziami (rys. **7.14**).

RYSUNEK 7.14. **PRZYKŁADOWA TABELA ZAŁADOWANA DO OPENREFINE**

Źródło: zrzut ekranu.

Następnie należy utworzyć projekt – przycisk „Create Project" (rys. **7.15**).

RYSUNEK 7.15. **NOWY PROJEKT W OPENREFINE**



Źródło: zrzut ekranu.

Sortowanie

Zacznijmy od definiowania rekordów w systemie. Pierwsza kolumna tabeli musi zawierać identyfikator rekordu. Wszystkie rekordy z pustą pierwszą kolumną muszą być powiązane z poprzednim rekordem o podobnym identyfikatorze. Wykorzystamy w tym celu sortowanie:

- przenosimy kolumnę identyfikatora na początek tabeli (Editcolumn>Movecolumn to beginning),
- sortujemy według pierwszej kolumny „Identyfikator" (Sort > by text),
- sortujemy według cechy (Sort > by text),
- zmieniamy kolejność na stałe (opcja w górnym panelu tabeli: Sort > Reorder rows permanently).

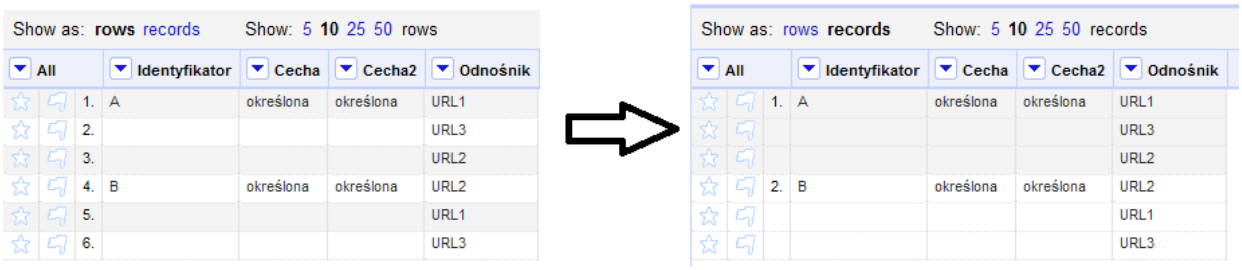
Po tych przekształceniach powinniśmy otrzymać następującą tabelę:

IDENTYFIKATOR	CECHA	CECHA2	ODNOŚNIK
A	określona	określona	URL1
A			URL3
A			URL2
B	określona	określona	URL2
B			URL1
B			URL3

Definicja rekordów

- Przełączamy „Show as" na „rows" i usuwamy zduplikowane identyfikatory (na kolumnie „Identyfikatory" wybieramy Edit cells > Blank down).
- Przełączamy „Show as" z powrotem na „records". Teraz rekordy z podobnymi identyfikatorami mają wspólny kolor (rys. **7.16**).

RYSUNEK 7.16. **GRUPOWANIE REKORDÓW Z PODOBNYMI IDENTYFIKATORAMI W OPENREFINE**



Źródło: zrzut ekranu.

Łączymy odnośniki.

- Łączymy komórki kolumny „Odnośnik" (Edit cells > Join multi-valued cells).
- W wyświetlonym oknie określamy separator (przecinek ze spacją).

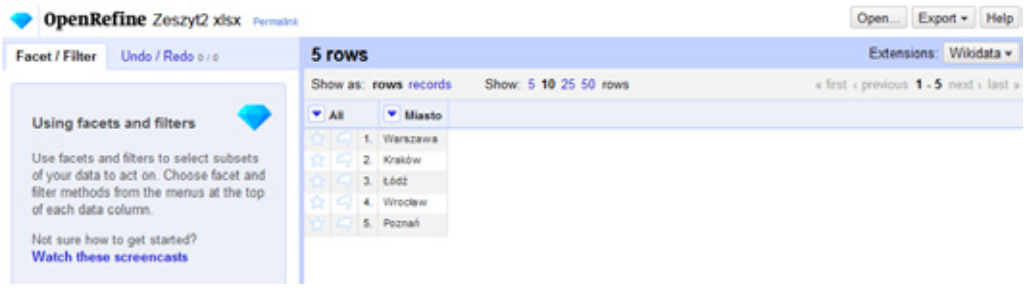
W wyniku otrzymamy poniższą tabelę:

IDENTYFIKATOR	CECHA	CECHA2	ODNOŚNIK
A	określona	określona	URL1, URL3, URL2
B	określona	określona	URL2, URL1, URL3

Przykład 2. Wzbogacenie danych

Rozpatrzmy przykład wzbogacenia danych przy pomocy Wikidanych w ramach OpenRefine. Dane wejściowe to tabela z jedną kolumną o nazwie „Miasto", w której wpisano 5 polskich miast (w każdej linijce jedno miasto): Warszawa, Kraków, Łódź, Wrocław, Poznań. Zatem mamy tylko 1 kolumnę oraz 6 wierszy (razem z nagłówkiem tabeli). Przykładowa tabela została zapisana w formacie Excel. Podobnie jak w poprzednim przykładzie, ładujemy ten plik do OpenRefine i tworzymy projekt (rys. **7.17**).

RYSUNEK 7.17. **TABELA WEJŚCIOWA PRZED WZBOGACENIEM DANYCH W OPENREFINE**



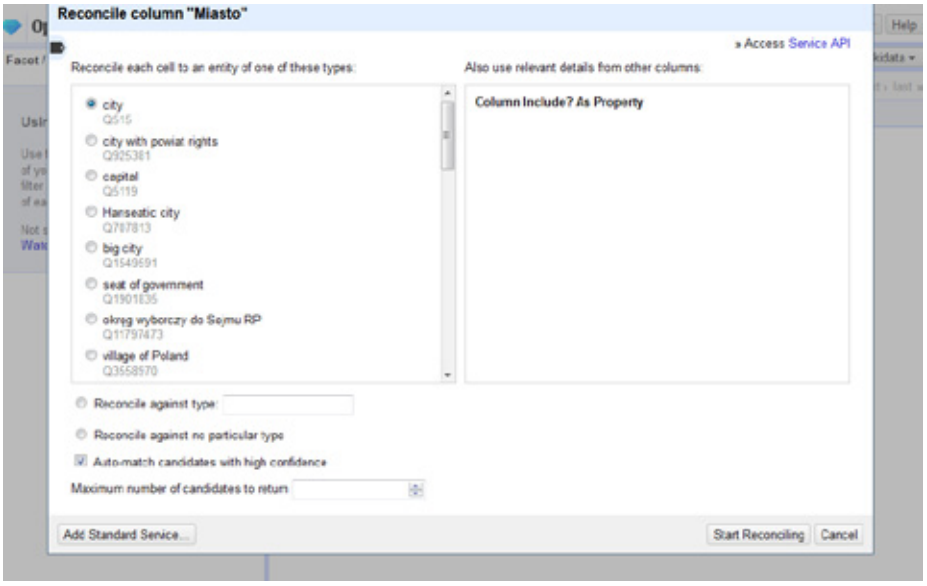
Źródło: zrzut ekranu.

Musimy teraz uzgodnić dane w naszej tabeli, co oznacza, że musimy przekazać dane o tym, co oznaczają wartości wpisane do poszczególnej kolumny. Zatem w naszym przypadku możemy to zrobić tylko dla jednej kolumny. W tym celu w kolumnie wybieramy „Reconcile" > „Start reconciling". W nowym oknie wybieramy „Wikidata (en)". Następnie musimy wybrać typ danych z listy proponowanych, w naszym przypadku najbardziej odpowiadającym jest „Q515 city" (rys. **7.18**), po tym należy wcisnąć „Start Reconciling".

Następnie dla kolumny „Miasto" należy wybrać „Edit columns" > „Add columns from reconciled values". W wyświetlonym oknie należy wybrać cechy, o które chcemy wzbogacić nasze dane. Na przykład: chcemy dla każdego miasta mieć dodatkowe kolumny z liczbą ludności (szukamy „population" z identyfikatorem P1082) oraz nazwiskiem prezydentem miasta (szukamy „head of government" z identyfikatorem P6). Po wyborze odpowiednich cech otrzymamy widok podobny do rys. **7.19**.

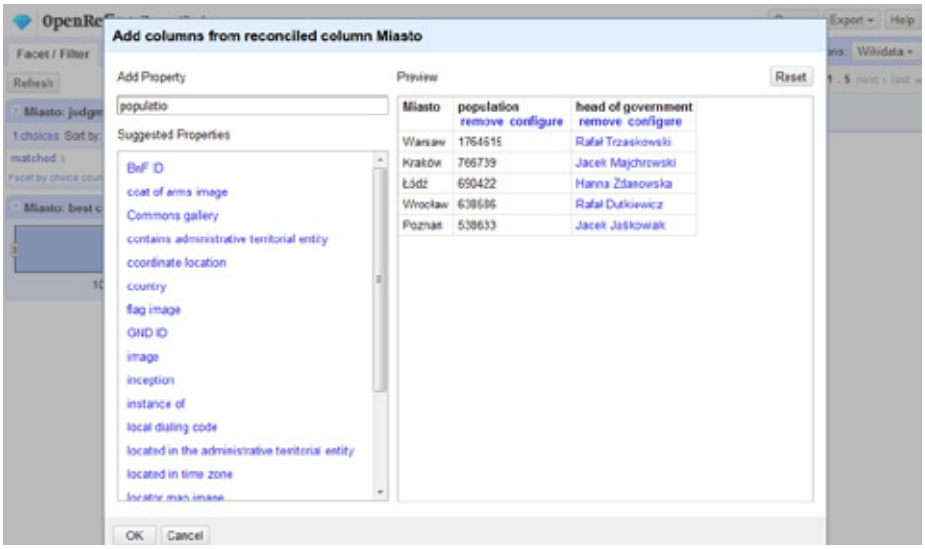
W wyniku otrzymamy tabelę przedstawioną na rys. **7.20** z 3 kolumnami zawierającymi dodatkowe informacje pochodzące z Wikidanych. Wynikową tabelę można zapisać w różnych formatach (w tym do formatu Excel, CSV, HTMl oraz innych.)

RYSUNEK 7.18. **LISTA CECH DO WYBORU DO KOLUMNY „MIASTO”**



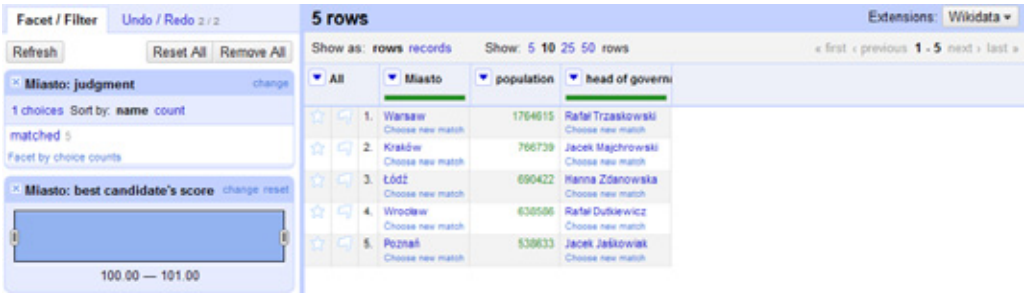
Źródło: zrzut ekranu.

RYSUNEK 7.19. **OKNO Z LISTĄ CECH DLA UTWORZENIA NOWYCH KOLUMN NA PODSTAWIE DANYCH Z KOLUMNY „MIASTO”**



Źródło: zrzut ekranu.

RYSUNEK 7.20. **TABELA O MIASTACH PO WZBOGACENIU O DANE Z WIKIDANYCH W OPENREFINE**



Źródło: zrzut ekranu.

7.4.3. WikiRank

Serwis internetowy WikiRank posiada interfejs API[23], który umożliwia pobieranie danych o ocenie jakości oraz popularności danych artykułów Wikipedii w formacie JSON. Na przykład na zapytanie o artykuł „Gniezno" w polskojęzycznej Wikipedii, dostaniemy odpowiedz pokazaną w listingu **7.1**.

KOD ŹRÓDŁOWY 7.1. **WYCINEK KODU ODPOWIEDZI INTERFEJSU WIKIRANK API NA ZAPYTANIE „GNIEZNO" W POLSKOJĘCYCZNEJ WIKIPEDII**

```
{
    "query":{
        "name":"Gniezno",
        "lang":"pl"
    },
    "result":{
        "pl":{
            "popularity":100,
            "name":"Gniezno",
            "quality":99.2173
        },
        "en":{
            "popularity":26.2692,
            "name":"Gniezno",
            "quality":54.4553
        },
    }
}
```

23 https://api.wikirank.net/

W celu utworzenia zapytania można korzystać z adresu URL wg schematu:

https://api.wikirank.net/api.php?name={nazwa artykułu}&lang={kod jezyka}

Dla przykładowego artykułu o Gnieźnie w polskojęzycznej Wikipedii adres URL będzie miał postać: https://api.wikirank.net/api.php?name=Gniezno&lang=pl

W celu automatyzacji procesu pobierania oraz przetwarzania otrzymanych wyników, rozważymy gotowy algorytm w języku Python pokazany w listingu **7.2**. Program ten pobiera dane z WikiRank API dla artykułu „Gniezno" z polskojęzycznej wersji Wikipedii (pl) i wyświetla 5 wersji językowych o najwyższej jakości oraz popularności.

KOD ŹRÓDŁOWY 7.2. **PROGRAM W JĘZYKU PYTHON DO POBIERANIA DANYCH Z WIKIRANK API**

```
import requests, json

url = 'https://api.wikirank.net/api.php'
dane = {"name": "Gniezno", "lang":"pl"}
r = requests.post(url, json=dane)
js = json.loads(r.text)

quality = {}
popularity = {}

for key,value in js["result"].items():
    quality[key] = value["quality"]
    popularity[key] = value["popularity"]

limit = 5

print("Wersje językowe o najwyższej jakości:")
iteracja = 0
for nazwa in sorted(quality, key=quality.get, reverse=True):
    iteracja += 1
    if iteracja>limit: break
    print(nazwa +" : "+str(quality[nazwa]))

print ("----------")
print ("Najpopularniejsze wersje językowe:")
iteracja = 0
for nazwa in sorted(popularity, key=popularity.get, reverse=True):
    iteracja += 1
    if iteracja>limit: break
    print(nazwa +" : "+str(popularity[nazwa]))
```

W związku z tym, że WikiRank API zwraca dane dla wszystkich dostępnych języków bez sortowania, algorytm wybiera tylko wersje językowe z najwyższą wartością jakości oraz popularności (odpowiednio wiersze 18–22 oraz 26–30 w listingu **7.2**). Dla zapytania o artykuł „Gniezno" w polskojęzycznej Wikipedii przy ograniczeniu do 5 wersji językowych (zmienna „limit" wierszu nr 15) otrzymamy odpowiedź przedstawioną na listingu **7.3**.

KOD ŹRÓDŁOWY 7.3. **WYNIK DZIAŁANIA KODU 7.2 DLA ARTYKUŁU O GNIEŹNIE**

```
1  Wersje językowe o najwyższej jakości:
2  pl : 99.2173
3  he : 65.7262
4  en : 54.4553
5  de : 48.7632
6  es : 35.7932
7  ----------
8  Najpopularniejsze wersje językowe:
9  pl : 100
10 en : 26.2692
11 de : 13.9911
12 ru : 13.1124
13 cs : 4.285
```

DBpedia

Dane w DBpedii pochodzą z różnych wersji językowych Wikipedii. Nazwy poszczególnych parametrów infoboksów są różne w różnych językach. Dzięki odwzorowaniom utrzymywanym w ramach DBpedii można te nazwy ujednolicić, tłumacząc je na własności z ontologii DBpedii. Daje to możliwość porównania wartości tego samego parametru infoboksów w różnych językach. Biorąc pod uwagę fakt, że infoboks jest umieszczany w artykułach, w celu porównania wersji językowych możemy zastosować niektóre z najbardziej istotnych miar jakości artykułów opisanych w sekcji **7.4.1.** Na przykład naszym celem jest wybór określonego parametru o najwyższej jakości z infoboksów opisujących PKN Orlen w 4 wersjach językowych: angielskiej (EN), niemieckiej (DE), polskiej (PL), rosyjskiej (RU), przedstawionych na rys. **7.21**.

Jednym z parametrów, którego wartość może się zmieniać w czasie, jest „zatrudnienie", które w ontologii DBpedii ma nazwę „numberOfEmployees". Biorąc pod uwagę opis algorytmu w WikiRank API (patrz sekcja **7.4.3**) oraz możliwości korzystania z serwera ekstraktorów DBpedii (patrz sekcja **4.6.4**) można napisać skrypt, który będzie wybierał wartość parametru z najlepszej wersji językowej.

RYSUNEK 7.21. **INFOBOKSY OPISUJĄCE PKN ORLEN W 4 WERSJACH JĘZYKOWYCH**

Źródło: opracowanie własne.

KOD ŹRÓDŁOWY 7.4. **PROGRAM W JĘZYKU PYTHON DO WYBORU NAJLEPSZEGO PARAMETRU Z ARTYKUŁÓW WYBRANYCH WERSJI JĘZYKOWYCH WIKIPEDII**

```
1   import requests, json
2   nazwa = "PKN Orlen"
3   jezyk = "pl"
4   parametr = "numberOfEmployees"
5   jezyki_do_porownania = ["en","de","ru","pl"]
6
7   dbpediapar = 'http://dbpedia.org/ontology/'+parametr
8   url = 'https://api.wikirank.net/api.php'
9   data = {"name": nazwa, "lang":jezyk}
10  r = requests.post(url, json=data)
11  js = json.loads(r.text)
12
13  quality = {}
14  jezyk_nazwa = {}
15  jezyk_parametr = {}
16
17  for key,value in js["result"].items():
18      if key not in jezyki_do_porownania: continue
19      quality[key]=2/((1/(value["quality"]))+1/(value["popularity"]))
20      jezyk_nazwa[key] = value["name"]
21
22      main_url = 'http://mappings.dbpedia.org/server/extraction/' + key + '/extract?title='
23      url = main_url + value["name"] + '&format=rdf-json&extractors=mappings'
24
25      r = requests.get(url)
26      for line in r.text.splitlines():
27          js = json.loads(line[:-1])
28          if dbpediapar in (js[list(js)[0]]):
29              jezyk_parametr[key]=js[list(js)[0]][dbpediapar][0]["value"]
30
31  najlepszawersja = sorted(quality, key=quality.get, reverse=True)[0]
32  print(parametr+" ("+najlepszawersja +"): "+str(jezyk_parametr[najlepszawersja]))
```

W wierszach nr 2-5 należy wpisać dane wejściowe:

- „nazwa" i „jezyk" wskazują na nazwę artykułu oraz wersji językowej, z której zaczynamy szukać inne wersje do porównania,
- „parametr" – unifikowana nazwa parametru, którego wartości należy porównać,
- „jezyki do porownania" – wersje językoweWikipedii, które należy wziąć pod uwagę przy porównaniu wartości.

Algorytm uwzględnia również różne nazwy artykułów w poszczególnych wersjach językowych Wikipedii. Przy zadanych zmiennych algorytm zwróci „numberOfEmployees (pl): 20262", co oznacza, że według polskojęzycznej Wikipedii liczba zatrudnionych w PKN Orlen wynosi 20.262 osób i te dane są najwyższej jakości spośród rozpatrywanych wersji językowych. Lista wartości parametru „numberOfEmployees" wraz z miarami syntetycznymi jakości oraz popularności w rozpatrywanych wersjach językowych Wikipedii dla artykułu o PKN Orlen przedstawione są w tabeli **7.14**.

TABELA 7.14 **WARTOŚCI PARAMETRU „NUMBEROFEMPLOYEES" ORAZ MIARY JEGO JAKOŚCI DLA ARTYKUŁU O PKN ORLEN**

JĘZYK	NUMBER OF EMPLOYEES	JAKOŚĆ	POPULARNOŚĆ
PL	**20262**	**60**	**100**
EN	24113	28	48
DE	19730	39	15
RU	24113	24	7

Źródło: opracowanie własne.

Należy zwrócić uwagę, że przedstawiony sposób nie jest idealnym rozwiązaniem problemu porównywania jakości danych pomiędzy wersjami językowymi Wikipedii, ponieważ była brana pod uwagę bardzo ograniczona liczba miar jakości. Zatem należy rozważyć uwzględnienie dodatkowych miar, w tym opisanych w niniejszym raporcie.

8 Rekomendacje i wnioski końcowe

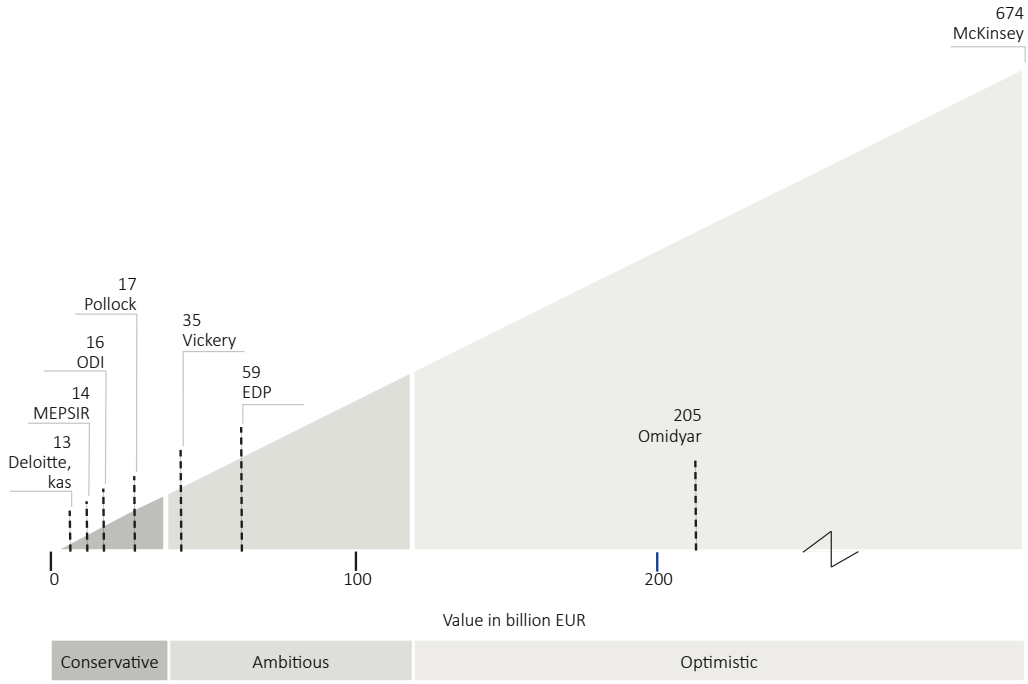
W niniejszym rozdziale dokonujemy podsumowania zagadnień poruszonych w raporcie oraz przedstawiamy całościowe wnioski i rekomendacje opracowane na podstawie przedstawionych treści. Punktem wyjścia była ocena dostępności danych otwartych dla przedsiębiorców w Polsce. W szczególności zwracamy uwagę na zasoby danych jako element kreowania wartości dla przedsiębiorców. Zaproponowana typologia źródeł może być wskazówką, gdzie oraz jak znaleźć dane o jakości niezbędnej do wykorzystania w biznesie. Na końcu przedstawiamy rekomendacje dla różnych podmiotów działających w obszarze danych otwartych. Zaprezentowane zostały także ogólne rekomendacje dotyczące obszarów, które pozwalają na wykorzystanie danych do budowania usług o wysokim stopniu innowacyjności i jakości – co przekłada się na zwiększenie konkurencyjności przedsiębiorstwa.

8.1. Zasoby danych jako element kreowania wartości

Analizując znaczenie tematyki zasobów danych poruszonych w raporcie, należy przede wszystkim podsumować ekonomiczne aspekty pozwalające na zwiększenie konkurencyjności polskich firm wynikające z opisanych zasobów. Konkurencyjność ta może być zapewniona nie tylko przez ulepszanie istniejących już usług, ale także przez możliwość powstania nowych usług w oparciu o dane (o charakterze innowacyjnym), zarówno w perspektywie usług komercyjnych, jak i tych świadczonych przez instytucje sektora publicznego.

Rola zasobów danych, rozumianych jako zbiory otwarte (open data) opisywane w tym raporcie, może być jednolita bez względu na źródło czy podmiot je wykorzystujący. Utrzymywanie otwartych zbiorów danych i ich wykorzystanie tworzy samo w sobie wartość dodaną [**13**] i sprzyja tworzeniu innowacji [**81**], a także ma bezpośrednie przełożenie na tworzenie usług przez przedsiębiorców [**204**]. To tworzenie wartości poprzez udostępnianie otwartych zbiorów można rozumieć zarówno jako tworzenie nowych usług w oparciu o nie (tak jak aplikacji wymienionych w raporcie), jak i otrzymywanie sprzężenia zwrotnego w postaci statystyk wykorzystania. Wartość zbioru danych czy też usługi udostępnionej na poziomie API można pośrednio badać wskazując na ich popularność wyrażaną jako liczba pobrań czy wywołań API. Wartość danych można przedstawiać jednak nie tylko opisowo, ale podać również konkretne liczby. Na przykład: Europejski Portal Danych (European Data Portal) w raporcie z grudnia 2017 [**19**] zebrał różne opracowania poświęcone szacowaniu wartości samego sektora otwartych danych, przygotowane przez najbardziej znane firmy analityczne. Wyniki tych opracowań zostały przedstawione w czytelnej formie ułatwiającej porównanie na rys. **8.1**. To, co jest od razu zauważalne, to bardzo duża rozbieżność oszacowań – wartość otwartych danych mieści się w zakresie między 13 a ponad 650 miliardów euro.

RYSUNEK 8.1. **OSZACOWANIA WARTOŚCI OPEN DATA**



Źródło: Raport European Data Portal [19]

Wartość ta jednak jest związana nie tylko ze zbiorami per se, ale ich wykorzystaniem w różnych scenariuszach. Z perspektywy dotychczasowych zastosowań otwartych zbiorów danych udostępnionych przez administrację (w rozumieniu usług opisanych w rozdziale **3**) jako najpopularniejsze należy wymienić następujące obszary [**19**, **80**, **110**, **132**, **150**][1]:

- Ochrona środowiska – poprzez budowanie usług analizujących stan obecny i aktywizację przedsiębiorców oraz obywateli w tym kierunku.
- Analiza mobilności i przestrzeni miejskiej oraz zagospodarowania przestrzennego – przez łączenie danych ze spisów publicznych wraz z danymi open data o charakterze geograficznym (dane z OpenStreetmap, serwisów WMS, zbiorami z telekomów i social media) i danymi z sensorów (natężenia ruchu, dźwięku, jakości powietrza), co w efekcie pozwoli otrzymać szacunki o mniejszym koszcie i dużo większej rozdzielczości i zakresie badanych zmiennych.
- Analiza działania usług o charakterze krytycznym – np. sieci energetyczne, transport.
- Administracja państwowa i aktywizacja obywateli, a także zwiększanie zakresu udostępnionych usług.

1 Także w perspektywie dyrektywy wskazanej w sekcji **3.3.2**.

Wskazując na potencjalne zyski z wykorzystania zbiorów danych, posłużyć się można podsumowaniem przedstawionym w tabeli **8.1**. Co ważne, jako istotny partner we wdrażaniu takich rozwiązań wskazywane są także jednostki naukowe [**141**]. Wnioski przedstawione w opracowaniu wskazują również na problemy z perspektyw różnych uczestników. Uwzględnienie rozwiązań systemowych niwelujących potencjalne zagrożenia może być kluczowe w przypadku opracowywania strategii wykorzystania zbiorów danych; możliwe, że nawet bardziej niż podkreślanie potencjalnych zysków.

TABELA 8.1. **PODSUMOWANIE SZANS I ZAGROŻEŃ WYKORZYSTANIA OTWARTYCH ZASOBÓW DANYCH DLA POSZCZEGÓLNYCH AKTORÓW**

UCZESTNIK	CELE	MOŻLIWE ZYSKI	POTENCJALNE ZAGROŻENIA
Przedsiębiorcy	Wykorzystanie badań naukowych w celu opracowywania produktów, usług i wdrożeń opartych na danych.	Innowacyjne produkty i przewaga konkurencyjna przez wykorzystanie nieeksploatowanych przez inne przedsiębiorstwa zbiorów danych. Optymalizacja procesów. Większa efektywność kosztowa i elastyczność zatrudnienia w R&D.	Niereplikowalność wyników lub wyniki niekorzystne (w formie opracowań lub nieudanych produktów).
Naukowcy	Rozpoznawalność w środowisku, możliwość publikacji artykułów w dobrych wydawnictwach.	Bezpośrednie przełożenie na karierę naukową, możliwość dodatkowego finansowania (badania zamawiane, współpraca w konsorcjach naukowo przemysłowych, doktoraty wdrożeniowe).	Przeciwstawne inicjatywy: publikacja i otwartość badań kontra chęć pozostawienia efektów jako tajemnicy handlowej przedsiębiorstwa. Niedopasowanie sytemu oceny badań naukowych do badań objętych tajemnicą handlową.
Organizacje	Wspieranie badań nie mających jasnego przełożenia na produkty, usługi i wdrożenia, które mogą przynieść bezpośredni zysk.	Wsparcie inicjatyw społecznych o wysokim poziomie wartości (które mogą być nieopłacalne na danym stadium badań).	Wykorzystanie kapitału bez widocznego efektu (badania naukowe) w przeciwieństwie do jasno aplikowalnych scenariuszy pomocy mających bezpośredni wpływ na społeczeństwo.
Administracja	Wsparcie dla obszarów nauki mających duże znaczenie dla gospodarki (bardziej efektywne niż bezpośrednie zlecanie badań).	Podnoszenie jakości kadry naukowej i transferu technologii pomiędzy przedsiębiorstwami i nauką; badania wspierające ważne sektory gospodarki.	Niski poziom kontroli nad wydatkowaniem środków (w szczególności, gdy opracowane rozwiązania są podmiotem tajemnicy handlowej i proponowanej przewagi konkurencyjnej).

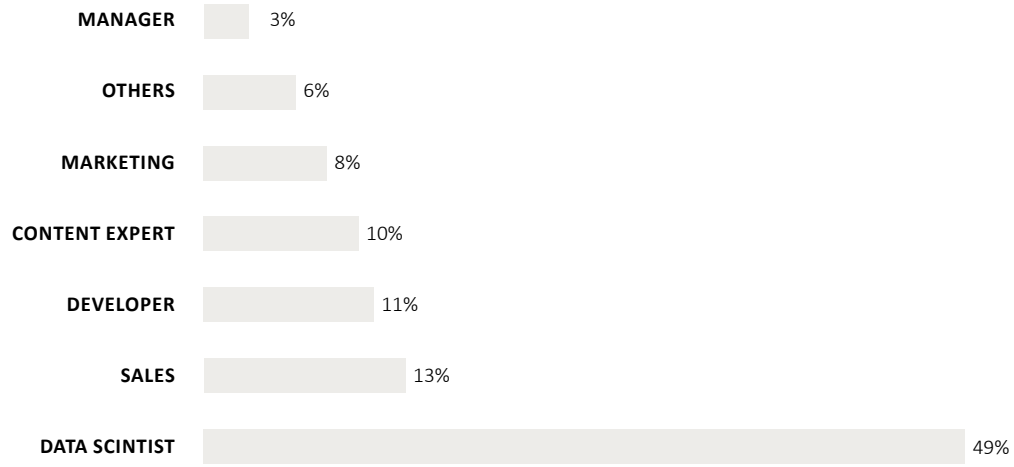
Źródło: opracowanie własne na podstawie [13, 79, 81, 108, 141, 204].

8.1.1. Rynek pracy i umiejętności w obszarze open data

Klasyfikując obszary zastosowań open data jako dziedziny nauki, należy wskazać przede wszystkim transport i logistykę, gospodarkę przestrzenną (także na pograniczu informatyki z uwzględnieniem tematyki smart cities), socjologię oraz wszystkie tradycyjne obszary analityczne, tj. statystykę czy ekonometrię. Efektywna analiza w tych obszarach wykorzystuje jednak metody operujące na dużych i zróżnicowanych zbiorach danych o charakterystyce zbiorów big data, co wymaga wykorzystania zupełnie innych technik analizy, jak uczenie maszynowe (machine learning), metody sztucznej inteligencji (rozumianej w większości jako głębokie uczenie sieci neuronowych), jak i wiedzy inżyniersko-informatycznej w kwestii efektywnego składowania i przetwarzania tych danych. Dodatkowo nowe typy przetwarzanych danych, tj. dane: tekstowe, o nieznanych rozkładach, pochodzące z sensorów, dot. procesów biznesowych czy stanu usług, o charakterze grafowym czy geograficznym na poziomie współrzędnych wymagają stanowczo innego zakresu umiejętności niż tradycyjne źródła statystyczne. Budowanie usług wykorzystujących wszystkie zalety takich zbiorów wymaga specjalistów o profilu badawczym odznaczającym się znajomością odpowiednich metod przetwarzania i specyficznych dla nich technologii, a także ich współpracy ze specjalistami z określonych dziedzin (np. w dziedzinie planowania przestrzennego).

Te wnioski stanowczo mają odbicie w profilu zatrudnienia wskazanym przez wyżej wymieniony raport [**19**]. Jak widać na rys. **8.2**, większość firm o profilu open data poszukuje specjalistów posiadających właśnie taki zakres umiejętności – najczęściej poszukiwane stanowisko to "Data scientist". To ogólne nazewnictwo opisuje specjalistę posiadającego wszystkie wyżej wymienione umiejętności, wraz ze zdolnością do samodzielnego prowadzenia prac badawczych (science) i rozwojowych. Oznacza to również, że możliwe jest zbudowanie zespołów w oparciu o kluczowe kompetencje wymienione wyżej – składające się na pełny profil umiejętności potrzebny do wdrożenia rozwiązań wykorzystujących open data.

RYSUNEK 8.2. **PREFERENCJE ZATRUDNIENIA W FIRMACH ZAJMUJĄCYCH SIĘ OPEN DATA**



Źródło: raport European Data Portal [19].

8.1.2. Modele biznesowe open data

Wdrażanie zastosowań opartych na open data wiąże się z pokonaniem wielu rodzajów wyzwań. Przedstawione w opracowaniu zaprezentowanym w tabeli **8.2** wyzwania wskazujące na potencjalne bariery we wdrażaniu rozwiązań opartych na danych zostały podzielone na aspekty: techniczne, prawne, ekonomiczne, organizacyjne i społeczne. Z perspektywy raportu najważniejszymi aspektami są kwestie: **spójności, możliwości odnalezienia zbiorów, struktury organizacyjnej i wsparcia technicznego**.

Dla przykładów zastosowań prezentowanych danych może powstać wiele modeli biznesowych pozwalających na kreowanie wartości dodanej. Korzystając z prostego diagramu, wskazującego na poziomy przetwarzania open data zaprezentowanym na rys. **8.2**., wskazać można poszczególne role w procesie wykorzystania danych:

- administrację publiczną,
- podmioty komercyjne publikujące dane,
- podmioty zajmujące się ekstrakcją i przetwarzaniem danych,
- podmioty zajmujące się analizą,
- podmioty skupione na dostarczaniu usług dla klienta końcowego (B2C).

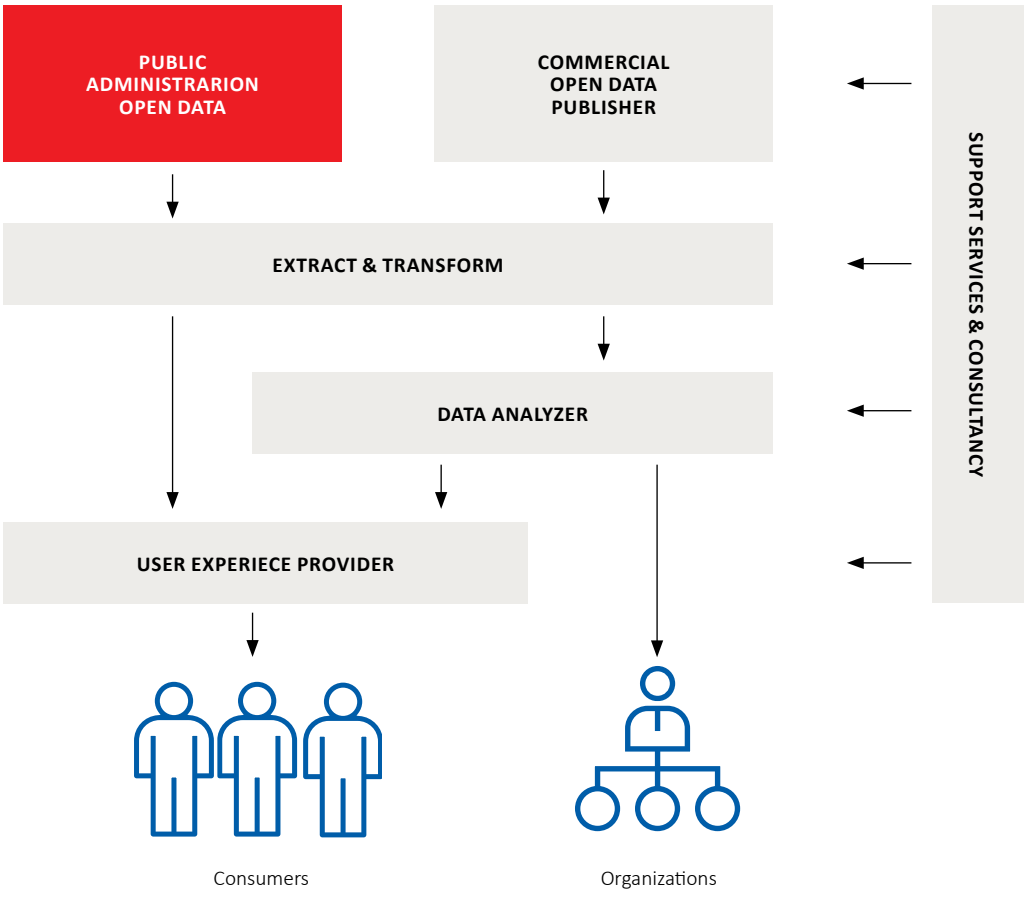
TABELA 8.2. **PROBLEMY I WYZWANIA ZWIĄZANE Z OPEN DATA**

TYP WYZWANIA	WYZWANIE	MOŻLIWE ROZWIĄZANIA
Techniczne	Format danych	Wykorzystanie formatów danych przetwarzalnych automatycznie (CSV, JSON)
	Niejednoznaczność	Używanie formatów deskryptywnych (dodatkowe informacje w pliku JSON opisujące zbiór); dodawanie dokumentacji i medatanych
	Możliwość odnalezienia	Korzystanie ze spójnych formatów metadanych i kategoryzacji; poprawienie wyszukiwarek; agregacja zasobów
	Spójność	Używanie spójnych formatów w wielu źródłach i tej samej reprezentacji podobnych zmiennych na różnych poziomach; Wykorzystywanie wersjonowania (np. za pomocą grafów zgodnych z RDF)
	Przepustowość	Dopasowanie się do standardów technicznych o odpowiedniej przepustowości; Testy obciążenia usług
Legislacyjne Prawne	Licencje	Spójna definicja polityk wykorzystania danych i adnotacja zbiorów
	Konflikty (niespójności)	Definicje spójnych polityk w sprawie zakresu, standardów i sposobu udostępniania danych na poziomie krajowym
	Regulacje prawne	Opracowywanie dokumentów mających na celu określenie zasad udostępniania danych
	Prywatność i ochrona danych	Definiowanie polityk prywatności i ochrony dla zbiorów danych; Ograniczanie dostępu z wykorzystaniem systemów kontroli dostępu
	Odpowiedzialność	Konsultacje społeczne; Podnoszenie świadomości; Określenie standardów technicznych i jakości dla publikowanych danych; Definicja dokumentów określających zakres odpowiedzialności za źródła danych
Ekonomiczne i Finansowe	Budżetowanie	Uwzględnianie budżetowania specjalnie dla projektów open data
Organizacyjne	Struktury organizacyjnej	Reorganizacja struktury organizacyjnej- wydzielenie zespołów odpowiedzialnych za utrzymanie open data; Definiowanie polityk i strategii dla zespołów skoncentrowanych wokół inicjatyw open data
	Pokrycia zakresów	Użycie spójnej architektury (np. portalu) i zakresu metadanych dla udostępnianych źródeł
	Wsparcia technicznego	Dostarczanie wsparcia dla jednostek adminstracji publicznej i klientów usług open data
Społeczne	Motywacyjne	Podnoszenie świadomości w temacie powtórnego wykorzystania open data i możliwych do osiągnięcia korzyści
	Dot. świadomości	Promowanie wartości open data wśród obywateli, poprzez promowanie przykładów wykorzystania i udanych inicjatyw
	Współudział obywateli	Tworzenie inicjatyw wskazujących na realne możliwości zastosowania open data dla inicjatyw obywatelskich (współtworzenia i wykorzystania)
	Konkurencyjne	Tworzenie konkursów i inicjatyw promujących wykorzystanie open data; Wprowadzanie kosztów za znaczące wykorzystanie udostępnianych zasobów (promujące wielu usługodawców zamiast monopolu ogranicza to jednak dostępność danych)

Źródło: Opracowanie własne na podstawie [14]

W zależności od umiejscowienia profilu firmy w powyższym procesie przetwarzania występują różnorodne modele biznesowe, które pozwalają budować wartość na podstawie zasobów danych. Przykładowe propozycje opisów modeli biznesowych, zgodnych z klasyfikacją przyjętą w powyższym podejściu, zostały zaproponowane w tabeli **8.3**. Co ważne, umieszczenie się w roli bardziej odległej w procesie, jako np. "Data Analyzer", wymaga danych, które zostały wcześniej przetworzone i przystosowane do potrzeb analizy. Zatem, o ile istnienie podmiotu na poziomie "Extract & Transform" może wydawać się nieopłacalne, to może on także świadczyć usługi dla podmiotów leżących dalej w procesie, zgodnie ze schematem na rys. **8.3**.

RYSUNEK 8.3. **PRZYKŁAD MODELU PROCESOWEGO KREOWANIA WARTOŚCI W OPARCIU O OPEN DATA**



Źródło: [108]

TABELA 8.3. **MODELE BIZNESOWE W OPARCIU O MODEL PROCESOWY KREOWANIA WARTOŚCI W OPEN DATA**

PROFIL MODELU BIZNESOWEGO	OFEROWANA WARTOŚĆ	MODEL FINANSOWANIA	ZASOBY	RELACJE Z PODMIOTAMI
Podmiot publikujący open data (Commercial Open Data publisher)	Publikacja (i produkcja) danych, które mogą być wykorzystane dalej jako open data	Zmniejszenie kosztów poprzez testowanie rozwiązań i interfejsów na poziomie udostępnianych usług	Utrzymywanie API dla udostępnianych danych	Społeczność developerów open data, którzy pozwalają zwiększyć świadomość użytkowników i wykorzystanie udostępnionych danych
Ekstrakcja i transformacja (Extract & transform)	Znajdowanie i konwersja danych na format pozwalający na dalsze przetwarzanie	Brak dobrze zdefiniowanego modelu biznesowego, organizacja działa pro-bono	open data i społeczność (wolontariusze-programiści, statystycy itd.) biorąca udział w procesie	Społeczność developerów open data i open source (w rozumieniu wymienionych wolontariuszy), administracja publiczna i podmioty komercyjne udostępniające dane
Analiza danych (Data Analyzer)	Tworzenie wizualizacji, analiza danych (oparta na algorytmach analitycznych i predykcyjnych) w celu generowania nowej wiedzy z udostępnionych danych	Projekty zamawiane, licencjonowanie własnego	open data, przetworzone i przygotowane dane (prywatne i komercyjne, także zescrapowane)	Społeczność developerów i użytkowników open data, podmioty administracji publicznej i międzynarodowe udostępniające dane (np. agencje europejskie)
Interfejs użytkownika (User experience provider)	Tworzenieużytecznych i innowacyjnych interfejsów dla użytkownika, które korzystają z danych open data	Reklamy, opłata za zakup produktu (jednorazowa), subskrypcja, licencjonowanie i model freemium	open data, przetworzone i przygotowane dane (prywatne i komercyjne, także zescrapowane)	Społeczność użytkowników open data i podwykonawcy (developerzy)
Pomoc i doradztwo (Support service and consultation)	Wsparcie innych przedsiębiorców w tworzeniu strategii i integracji (na poziomie technicznym, organizacyjnym) tak by stworzyć wartość dodaną sieci przedsiębiorstw	Projekty zamawiane oraz płatność za wykonane usługi i analizy	Zespoły programistycz-no-analityczne, podmioty lub dział wewnętrz-ny świadczący doradztwo i tworzący kontrakty	Społeczność developerów i użytkowników open data, małe przedsiębiorstwa chcące tworzyć rozwiązania open data (jako partnerzy wnoszący innowacyjne usługi dla dużych przedsiębiorców)

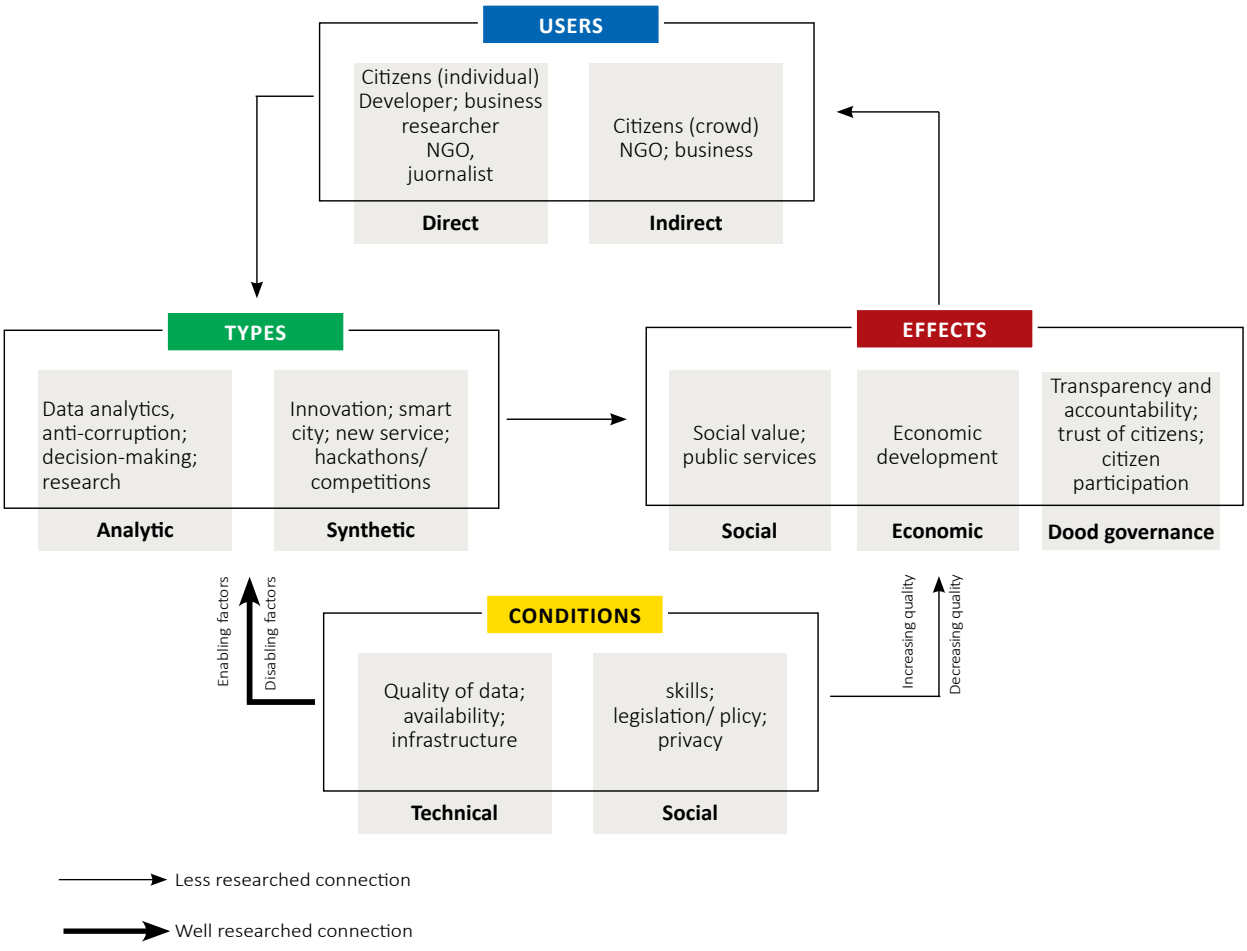
Źródło: tłumaczenie na podstawie [108].

W literaturze zaproponowano też całościowy model, który wskazuje na poszczególne obszary i typy zastosowania open data [**150**]. W oparciu o ponad 100 artykułów opisujących i kategoryzujących zastosowania open data autorzy zaproponowali model opisujący:

- **typy wykorzystania** open data – scenariusze wykorzystania: analityczny (deskryptywny względem procesów zachodzących i wspomagający podejmowania decyzji) i syntetyczny ("synthetic" – mający na celu projektowanie i tworzenie nowych rozwiązań możliwych do wdrożenia);
- **efekty wykorzystania** w podziale na: społeczne (usługi publiczne i szeroko rozumiane efekty społeczne), ekonomiczne (rozwój w znaczeniu ekonomii, z wyszczególnieniem poziomu technologii i innowacyjności usług), podniesienie jakości administracji (governance, przejrzystość, zaufanie, udział obywateli);
- **warunki**: techniczne (jakość danych, dostępność, infrastruktura) i społeczno-administracyjne (umiejętności, otoczenie prawne i społeczne, kwestie prywatności).

Przedstawiony przez autorów model prezentuje całościowe podejście, identyfikujące aktorów, typy wykorzystania i potencjalne efekty, a także warunki funkcjonowania wdrażanych usług (związane z szeroko pojętym otoczeniem gospodarczym). Wnioski z powyższych badań pozwalają na stworzenie własnego modelu biznesowego, przy uwzględnieniu akceptacji aktualnego poziomu wyzwań obecnych w obszarze open data (tabela **8.2**). Zaprezentowany model można zobaczyć na rysunku **8.4**. Jako że kreowanie wartości przez usługi open data wspomaga wszystkich aktorów, a największym producentem jest administracja publiczna, kluczowym więc z perspektywy administracji wydaje się także zwrócenie uwagi na przydatność i możliwość budowania przewagi konkurencyjnej i innowacyjnych usług w oparciu o publikowane przez nią dane. Zaproponowany **model może być dobrą rekomendacją dla przeprowadzenia procesu oceny i klasyfikacji przykładów usług zbudowanych głównie w oparciu o wykorzystanie danych.** Całościowa ocena wartości w różnorodnych obszarach zaproponowana w modelu może posłużyć jako wzorzec do oceny nowych proponowanych rozwiązań czy projektów mających na celu uzyskanie finansowania w obszarze wykorzystania zbiorów danych (open data) do kreowania wartości.

RYSUNEK 8.4. **MODEL OPISUJĄCY CAŁOKSZTAŁT ZASTOSOWANIA OPEN DATA W LITERATURZE NAUKOWEJ I PRZYKŁADACH BIZNESOWYCH**



Źródło: [150]

8.2. Rekomendacje związane z cyklem życia danych

Poniżej przedstawiamy wnioski z raportu i rekomendacje pogrupowane według cyklu życia, tj. od samego źródła do wykorzystania danych.

8.2.1. Typologia źródeł – ocena jakości danych

Źródła przedstawione w raporcie można oceniać za pomocą różnorodnych kryteriów, które zależą również od tego, z jakim ocenianym obiektem mamy do czynienia. W trakcie analizy przedstawionej w raporcie wyróżniliśmy kilka obiektów, które mogą zostać poddane analizie: **a) serwisy API udostępniające dane**, **b) agregatory i katalogi danych** oraz **c) statyczne zbiory danych**.

Mówiąc o ogólnej typologii dla wszystkich tych trzech kategorii, można oceniać jakość tych usług mówiąc o takich charakterystykach, jak:

- rzetelność źródła: rządowe, naukowe, komercyjne, prywatne,
- poziom odpowiedzialności za jakość dostarczanych danych,
- licencja,
- wielkość zbioru, istotna w szczególności dla zastosowań wykorzystujących uczenie maszynowe,
- aktualność źródła: dynamiczne (API), publikowane z opóźnieniem, publikowane regularnie, nieregularnie, nieaktualne,
- format, w tym jego wewnętrzna spójność i zgodność z przyjętymi standardami,
- poziom opisu metadanych,
- efektywność techniczna rozwiązania,
- popularność zbioru czy usługi,
- koszty, znaczące w szczególności w zastosowaniach takich, jak zbiory Google Big Query,
- przystępność usługi, w rozumieniu wymaganej wiedzy technicznej do jego wykorzystania.

W przypadku statycznych zbiorów można je analizować pod kątem takich cech jak: wielkość zbioru, kompletność, precyzja, ziarnistość, rozdzielczość. Przykładem mogą być katalogi wskazane przez strony takie, jak UCL Machine Learning repository[2] lub też przykład opisu zaprezentowanego przez Wikipedię (podrozdział **6.5.1**).

Mówiąc o **metrykach, które pozwalałyby jednoznacznie skwantyfikować jakość omawianych zbiorów i źródeł danych w sposób obiektywny**, warto się posłużyć modelem miar zastosowanym w bieżącej literaturze naukowej [**181**]. Przedstawiony tam model stawia sobie za cel zdefiniowanie miar, które są: mierzalne (możliwe do wyrażenia w reprezentacji

2 https://archive.ics.uci.edu/ml/

liczbowej), interpretowalne, możliwe do zagregowania od poziomu rekordu do zbioru, możliwe do wyliczenia w sposób automatyczny. Warto przy okazji zauważyć, że cechy te są mocno związane z przedstawionymi w podrozdziale **2.1** cechami informacji oraz technicznej zgodności ze standardami. Opis cech, które można zastosować do oceny zbioru danych, znajdują się w tabeli **8.4**.

TABELA 8.4. **MIARY JAKOŚCI OTWARTYCH ZBIORÓW DANYCH**

WYMIAR	MIARA	POZIOM	OPIS
Identyfikowalność	Informacja o utworzeniu	Zbiór danych	Wskazuje obecność lub brak metadanych związanych z procesem tworzenia zestawu danych.
Identyfikowalność	Informacja o aktualizacji	Zbiór danych	Wskazuje istnienie lub brak metadanych związanych z aktualizacjami wykonanymi w zestawie danych.
Aktualność	Procent aktualnych wierszy	Komórka	Wskazuje procent wierszy zestawu danych, które mają bieżące wartości, oznacza to, że nie mają żadnej wartości odnoszącej się do poprzedniego lub kolejnego okresu.
Aktualność	Opóźnienie w publikacji	Zbiór danych	Wskazuje stosunek między opóźnieniem w publikacji (liczbą dni, które minęły między momentem, w którym informacje są dostępne, a publikacją zbioru danych) i czasem określonym przez zbiór danych (tydzień, miesiąc, rok).
Aktualność	Opóźnienie po wygaśnięciu	Zbiór danych	Wskazuje stosunek opóźnienia w publikacji zbioru danych po wygaśnięciu jego poprzedniej wersji i okresu czasu określonego przez zbiór danych (tydzień, miesiąc, rok).
Kompletność	Procent kompletnych komórek	Komórka	Wskazuje procent kompletnych komórek w zbiorze danych. Oznacza to komórki, które nie są puste i mają przypisaną znaczącą wartość (tj. wartość spójną z domeną całej kolumny).
Kompletność	Procent pełnych rzędów	Komórka	Wskazuje procent kompletnych wierszy w zbiorze danych. Oznacza to wiersze, które nie mają żadnej niekompletnej komórki.
Czytelność	Procent znormalizowanych kolumn	Komórka	Wskazuje procent znormalizowanych kolumn w zbiorze danych. Po prostu bierze pod uwagę kolumny reprezentujące pewne informacje, które mają powiązane z tym standardy (tj. Informacjegeograficzne).
Czytelność	eGMS	Zbiór danych	Wskazuje stopień, w jakim zbiór danych jest zgodny ze standardem e-GMS (w odniesieniu do podstawowych elementów zasadniczo sprowadza się do specyfikacji Dublin Core)
Czytelność	5-gwiazd	Zbiór danych	Wskazuje poziom w 5-gwiazdkowym modelu otwartych danych, w którym znajduje się zestaw danych i przewaga oferowana z tego powodu.

WYMIAR	MIARA	POZIOM	OPIS
Zrozumiałość	Procent kolumn z metadanymi	Komórka	Wskazuje procent kolumn w zestawie danych, który zawierał opisowe metadane. Te metadane są ważne, ponieważ umożliwiają łatwe zrozumienie informacji o danych i ich reprezentacji.
Zrozumiałość	Procent kolumn w zrozumiałym formacie	Komórka	Wskazuje procent kolumn w zestawie danych, który jest reprezentowany w formacie, który może być łatwo zrozumiany przez użytkowników, a także jest odczytywalny maszynowo.
Dokładność	Procent dokładnych komórek	Komórka	Wskazuje procent komórek w zestawie danych, który ma poprawne wartości zgodnie z domeną i typem informacji zbioru danych.
Dokładność	Dokładność w agregacji	Komórka	Wskazuje stosunek między błędem w agregacji i skalą reprezentacji danych. Ta metryka ma zastosowanie tylko do zestawów danych, które mają kolumny agregacji lub gdy istnieją dwa lub więcej zestawów danych odnoszących się do tych samych informacji, ale na innym poziomie szczegółowości.

Źródło: opracowanie własne oraz [181].

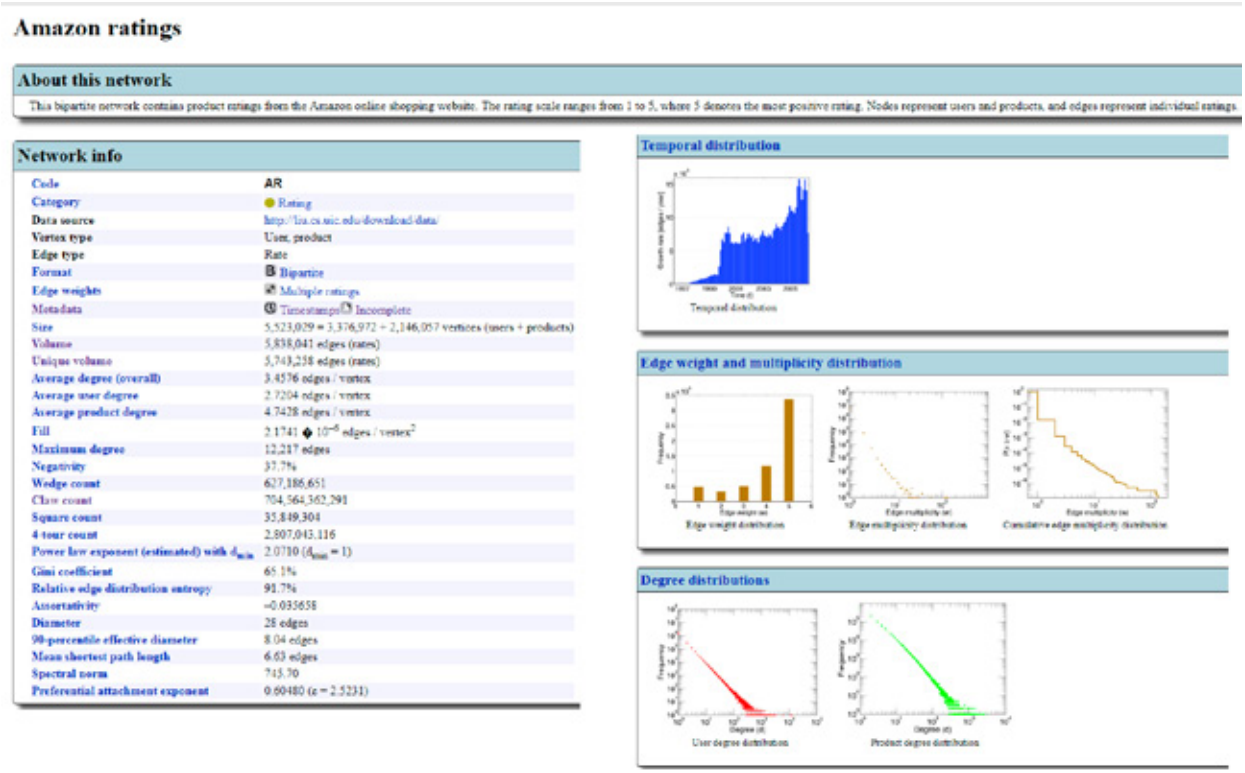
Jeżeli chodzi o 5-gwiazdkowy system oceny jakości otwartych danych, został on opisany w podrozdziale **2.2**.

Korzystanie z takiego modelu oceny jakości nie zwalnia jednak z walidacji wyników uzyskanych na ich podstawie, opisanych w procesie "nasycania danymi" w tym raporcie. W przypadku natomiast, gdy potrzebne jest rozszerzenie tych miar o bardziej specyficzne, pozwalające na odniesienie do charakterystyk zbiorów wyrażonych w postaci tekstu, obrazu, treści o charakterze społecznościowym, można posłużyć się procesem definicji dodatkowych atrybutów na podstawie artykułów w Wikipedii opisanych w tabeli **7.4**.

Charakterystyki mogą również różnić się w zależności od typu zbioru (geograficzny, medyczny, statystyczny, o charakterze grafu). Przykładowo duży zbiór grafowy Amazon Rating można analizować na poziomie charakterystyk grafu (rys. **8.5**). W przypadku badań wykorzystujących metody uczenia maszynowego warto wspomnieć także o reprezentatywności zbioru[3].

3 Można chociażby zwrócić uwagę na opracowania udostępnione na stronie https://www.fast.ai/. Z wprowadzeniem do tego tematu można się zapoznać w artykule dostępnym pod adresem https:// www.fast.ai/2019/01/29/five-scary-things/#bias w sekcji zatytułowanej *"AI encodes & magnifies bias"*.

RYSUNEK 8.5. **OPIS GRAFU DANYCH DLA ZBIORU "AMAZON RATING"**



Źródło: http://konect.uni-koblenz.de/networks/amazon-ratings

8.2.2. Pozyskiwanie danych

Struktura raportu wynika z podziału według najważniejszych typów źródeł, z których można pozyskiwać dane. Poniżej przedstawiamy pewne wnioski co do doboru źródeł – odpowiadając na pytanie "**skąd czerpać dane?**".

Mimo bardzo dużych wysiłków i rozwiązań również w prawodawstwie, czy to krajowym czy międzynarodowym, dane ciągle nie są udostępniane przez rządy w stopniu zadowalającym. Przedsiębiorstwa poszukujące danych z tej kategorii nie mają łatwego zadania. W oparciu o badania Barometru Otwartych Danych jako główny problem wskazano paradoksalnie, że tylko 1 na 10 zbiorów rządowych jest otwarty [**188**]. Po swoistym boomie mniej więcej 5 lat temu otwartość i dostępność danych nawet zmalała. Kolejnym czynnikiem do wzięcia pod uwagę jest jakość danych – dane rządowe są często niekompletne i słabej jakości. Portale danych wprawdzie są wdrożone, ale proces ich aktualizacji jest ręczny – co potwierdziliśmy

także podczas analizy publicznych zbiorów danych w tym raporcie. Dokładniejsze lub nowsze dane często można znaleźć w źródłach utrzymywanych przez dedykowane agencje rządowe, czyli poza oficjalnym rządowym portalem otwartych danych. Szczególnie widocznym przypadkiem są urzędy statystyczne – więcej danych referencyjnych można znaleźć właśnie tam, a nie na portalach rządowych. Jednak zakres danych statystycznych nie pokrywa wielu istotnych obszarów interesujących przedsiębiorców. Zauważalna jest również tendencja do publikowania jakichkolwiek danych, aby dobrze wyglądały w statystykach publikowanej witryny, natomiast nie są to dane, które mogą realnie wesprzeć rozwój przedsiębiorstw. Nie są to również dane, których potrzebują ludzie.

Bardzo dobrym rozwiązaniem może być korzystanie ze specjalnie opracowanych kolekcji (tzw. *curated list*), w celu identyfikacji użytecznych zbiorów danych. Reprezentatywnym przykładem jest OpenML opisany w sekcji **6.5.2**. Główne jego zalety to bardzo dużo zbiorów (rzędu 20 tys.), dobrze opisane, generowane automatycznie statystyki i podsumowania, wygodny interfejs do wyszukiwania z wieloma kryteriami, a także wgląd w ocenę społeczną prezentowanych zbiorów. Należy również uwzględnić to, że serwis ten łatwo integruje się z narzędziami oraz środowiskami analitycznymi, przez co nie jest konieczne ręczne ściąganie zbiorów, a jedynie umieszczenie odpowiedniego fragmentu kodu w programie.

8.2.3. Wykorzystanie open innovation

Główne założenie i cel wdrażania podejścia open innovation to aktywne wykorzystanie wiedzy znajdującej się poza przedsiębiorstwem oraz integracja jej ze stosowanym modelem biznesowym, a tym samym – wewnętrznymi zasobami. Pozyskując innowacje ze źródeł zewnętrznych istotną kwestią jest zarządzanie prawami własności intelektualnej [**142**]. Kwestie przekazania praw do komercyjnego wykorzystania opracowanych rozwiązań i zasobów, określenie wyłączności, postanowień licencyjnych czy wynagrodzenia są kluczowe dla powodzenia wykorzystanie otwartych innowacji w praktyce.

Podstawowe modele otwartych innowacji są następujące [**75**]:

- crowdsourcing,
- platformy produktowe (ang. *product platform*),
- innowacyjne sieci współpracy (ang. *collaborative innovative network*),
- konkursy innowacyjności (ang. *innovative contest*).

W praktyce najczęściej spotyka się modele mieszane, które łączą wszystkie (bądź większość) wyżej przedstawionych rozwiązań: zrzeszają społeczności innowacyjne, dostarczają zasobów danych i narzędzi na potrzeby użytkowników, jak również dostarczają rozwiązania techniczne umożliwiające przeprowadzanie konkursów innowacyjnych. W przypadku każdego z tych

modeli podstawą działania społeczności (użytkowników, uczestników) są zasoby – jeżeli są to zasoby organizacji, wówczas konieczne jest wyraźne określenie warunków ich wykorzystania.

Ponadto, aby możliwe było wykorzystanie opracowanej innowacji w praktyce i zintegrowanie jej z zasobami i know-how przedsiębiorstwa, konieczne jest określenie warunków przekazania rozwiązania – kwestia ta jest szczególnie istotna w przypadku konkursów innowacyjności.

Przedsiębiorstwa mogą występować w różnych rolach w procesach związanych z open innovation i tym samym, czerpać różne korzyści [**73**]. Po pierwsze, firmy mogą być beneficjentami, tj. jednostkami, dla których opracowywane jest innowacyjne rozwiązania. Występując w tej roli firma ma możliwość pozyskania rozwiązania przy pomocy zewnętrznych zasobów (przede wszystkim zewnętrznych ekspertów), wykorzystując równocześnie efekty społecznościowe – dostęp do ekspertów spoza organizacji, niezależnie pracujących nad rozwiązaniem problemu.

Beneficjenci czasami korzystają z pomocy innej firmy-organizatora, a czasami sami wcielają się w tę rolę, organizując proces pozyskania innowacji (zwykle konkurs) we własnym zakresie. Korzyści dla organizatora to przede wszystkim możliwość budowania społeczności, skupionej wokół określonego fenomenu (tematyki, organizacji, narzędzia) i możliwość włączenia takich czynności do modelu biznesowego (a co za tym idzie, korzyści finansowe z organizacji open innovation).

Firmy biorą udział w procesach innowacyjnych jako dostawcy zasobów. Jeżeli są to zasoby finansowe, przeznaczone na nagrody (wynagrodzenia) dla uczestników konkursów, wówczas są to najczęściej sami beneficjenci. Dostawcy zasobów mogą jednak zapewnić dane lub narzędzia do analizy. Takie działania pozwala dostawcy budować społeczność wokół własnych zasobów (np. specjalistycznego oprogramowania) oraz wspierać edukację w tym zakresie wśród użytkowników.

Ostatnią rolą są uczestnicy konkursów, a więc społeczność, która generuje rozwiązania innowacyjne na potrzeby beneficjenta. Uczestnikami są zarówno osoby prywatne, jak i firmy. Najważniejsze korzyści to przede wszystkim: zdobywanie doświadczenia, rozwijanie kompetencji w danym obszarze badawczym, a także budowanie wizerunku czy nawiązywanie kontaktów.

Firma, która chciałaby rozpocząć wykorzystanie otwartych innowacji, powinna przede wszystkim dokładnie zaplanować co i w jakim zakresie będzie outsource'owane do społeczności, jaki jest pożądany rezultat (jego format, zakres) oraz jak będzie zintegrowany z wewnętrznymi procesami B+R (np. w zakresie ustalenia warunków współpracy, sposobów wykorzystania rezultatów, rozstrzygnięcia kwestii własności intelektualnej). Konieczne jest również dotarcie do

wybranej grupy docelowej uczestników, a więc precyzyjne określenie wymagań, zasad oraz adekwatnej gratyfikacji.

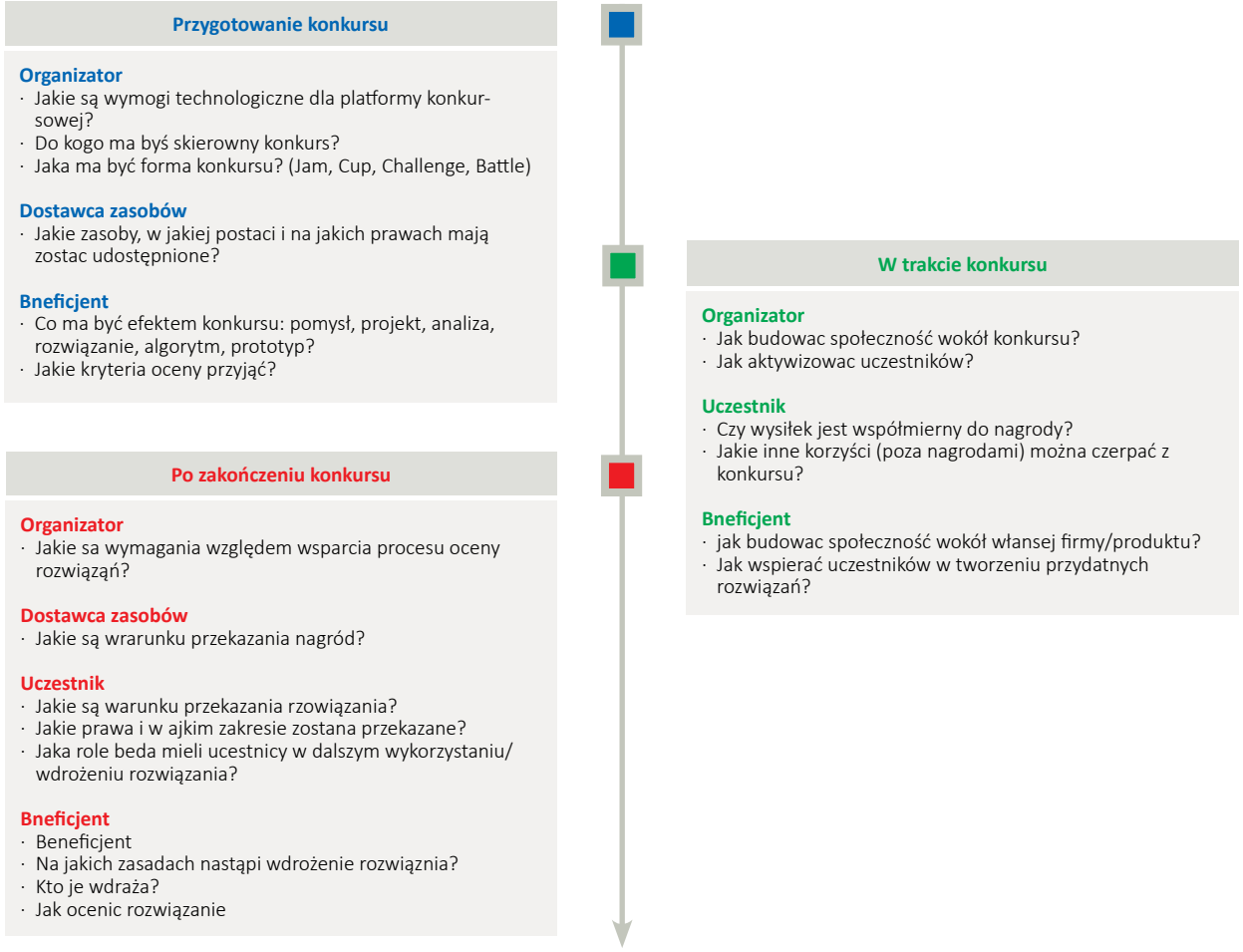
Jako, że najczęstszym zaobserwowanym w analizie przypadków modelem open innovation w Polsce były konkursy innowacyjności, to dla nich przygotowaliśmy dodatkowe rekomendacje. Podsumowując etapy procesu związanego z wdrażaniem open innovation, w szczególności uwzględniając proces organizacji konkursów inowacyjności opisany w sekcji **5.5.2**, opracowano krótki zestaw rad uwzględniający często pomijane elementy, zauważone podczas analizy przypadków przeprowadzonej w raporcie. Rekomendacje dla poszczególnych etapów procesu, zostały zaprezentowane na rysunku **8.6**. Jest to podsumowanie najważniejszych elementów, które należy uwzględnić przy organizacji takich wydarzeń. Dla organizacji zainteresowanych bardziej szczegółową analizą tematu open innovation, poleca się zapoznać z całym rozdziałem tego raportu poświęconym tej tematyce, a w szczególności z wymienioną wyżej sekcją.

8.2.4. Przechowywania danych

Istotnym problemem, który zauważyliśmy w raporcie, jest utrzymanie aktualności repozytoriów danych. Trafiliśmy na wiele ciekawych zbiorów, katalogów i list, które zawierały niestety nieaktualne linki. W efekcie nie można było odnaleźć wskazanych zbiorów lub też strony w ogóle były nieczynne. Z drugiej jednak strony utrzymywanie aktualności często jest nieuzasadnione ekonomicznie. Dotyczy to w szczególności tych repozytoriów, które były opracowane w ramach finansowanych projektów, dla których zostały rozwinięte zaawansowane interfejsy, a po zakończeniu finansowania nie było podmiotu, który przejąłby koszty utrzymania. Może to wskazywać na konieczność opracowania rekomendacji dla fundatorów badań naukowych w zakresie archiwizacji zbiorów danych zebranych podczas projektów. Nie tylko repozytoria nie są aktualne, ale czasami trudno wskazać, kto kogo importuje, gdzie jest źródło pierwotne, na ile aktualne i jak często aktualizowane są linki.

Z punktu widzenia konsumentów danych najlepiej uwagę poświęcić największym, najlepiej utrzymanym, najbardziej przydatnym zbiorom. Za takim podejściem przemawia charakterystyka publikacji zbiorów danych w ramach open science. Niewielkie zbiory (do 20 MB) najczęściej nie są w ogóle publikowane – 42% odpowiedzi naukowców. Wśród zbiorów o rozmiarze przekraczającym 50GB aż 70% jest współdzielonych, przy czym jest wtedy większa skłonność do korzystania z dedykowanych repozytoriów (59%) [**165**].

RYSUNEK 8.6. **RADY DOTYCZĄCE POSZCZEGÓLNYCH ETAPÓW ORGANIZACJI KONKURSU INNOWACYJNOŚCI**



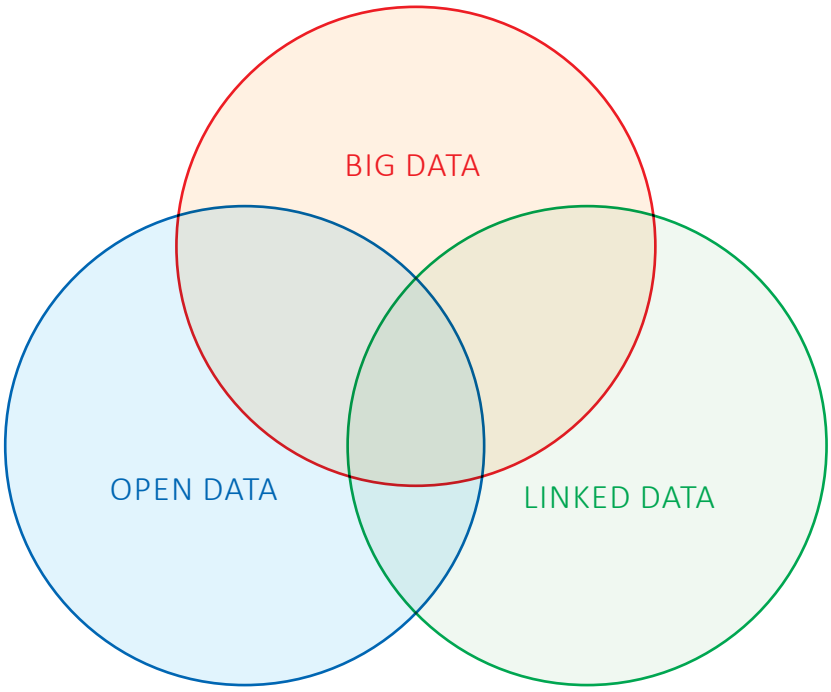
Źródło: opracowanie własne z wykorzystaniem serwisu draw.io.

Sytuacja w obszarze zbiorów przypomina tę z początków WWW – wtedy podejmowano wysiłki katalogowania stron, co niespecjalnie się sprawdzało biorąc pod uwagę szybkość przyrostu liczby stron. Teraz jednak nie ma wyraźnych dużych graczy, którzy skupialiby się na katalogowaniu zbiorów danych. Zgodnie z osiągnięciami Web 2.0 inicjatywę przejmują raczej pojedyncze osoby lub niezależne społeczności. Katalogi były nieaktualne i obecne „zbiory zbiorów" również takie są.

Uważamy również, że w szczególności dla przedsiębiorców dużo istotniejszy jest dostęp do danych dynamicznych. Oznacza to, że z punktu widzenia przedsiębiorcy cenniejsze jest źródło, które udostępnia dane bieżące przez API niż źródło, które udostępnia dane archiwalne w postaci pliku. Przedsiębiorcy powinni nie tylko utrzymywać dostęp do danych, ale również konserwować swoje dane poprzez dołączanie danych skojarzonych. Do tego celu doskonale nadanie się koncepcja „linked data". W naszym przypadku, jeśli chodzi o składowanie danych, rekomendacją byłoby zastosowanie podejścia big+open+linked data (rys. **8.7**):

- **big** – posiadamy dostęp do ogromnych zbiorów danych, możliwie szeroko charakteryzujących interesujące nas zjawisko,
- **open** – dane te są otwarte, mamy do nich swobodny dostęp, bez ograniczeń licencyjnych,
- **linked** – dane są udostępniane w postaci semantycznej, tzn. są same się opisują z wykorzystaniem globalnych identyfikatorów, które umożliwiają łączenie ze sobą wielu zbiorów bez kosztownych operacji integracji danych.

RYSUNEK 8.7. **PRZECIĘCIE BIG, OPEN I LINKED DATA**



Źródło: opracowanie własne.

8.3. Rekomendacje dla poszczególnych uczestników rynku

W perspektywie zasobów danych wskazać można kilka zasadniczych kwestii związanych z rekomendacjami w tym obszarze. Dla czytelności zostaną one podzielone jako rekomendacje dla: administracji rządowej, samorządowej i przedsiębiorców.

8.3.1. Rekomendacje dla administracji

- **Rozszerzenie serwisu dane.gov.pl o katalog API i usług udostępnianych dynamicznie.** Mówiąc o usłudze dane.gov.pl, wydawałoby się zasadne dla powyższego serwisu uzupełnienie go o katalog otwartych usług, opisanych w tym raporcie, takich jak – wyszukiwarka patentów, KRS, CEIDG. Dzięki temu platforma mogłaby pełnić rolę agregatora wszystkich usług udostępnionych przedsiębiorcom, organizacjom i obywatelom. Jest to zgodne z założeniami programu Polska Cyfrowa (patrz rozdział **3**). Warto także wspomnieć o rozszerzeniu katalogu dostępnych usług o niskim poziomie agregacji, dotyczących obszarów innych niż czysto demograficzne, takich jak: zdrowie, przestępczość, rynek pracy.
- **Budowa spójnej platformy dla udostępniania usług open data ze szczególnym uwzględnieniem API.** Przykład platformy na której zbudowane są usługi WMS, gdzie samorządy lokalne mogą wykorzystywać działającą usługę na poziomie państwowym do umieszczania w niej danych o charakterze lokalnym (np. plany zagospodarowania przestrzennego) może być najlepszym sposobem wdrażania innowacyjnych usług open data na poziomie lokalnym. Dobrym początkiem dla tego działania jest spójny standard API opublikowany przez Ministerstwo Cyfryzacji [**121**]. Stanowczo należy położyć nacisk na większą współpracę i celowość działań w zakresie analizy posiadanych danych **także na szczeblu państwowym.** Wiele bardzo ciekawych usług wymienionych w raporcie (np. widok.gov.pl, sekcja **3.2.2**), które są inicjatywami bardzo ciekawymi z perspektywy cyfryzacji samej administracji, wydaje się nie być odpowiednio popularna z powodu braku współpracy i spójnej strategii działania.
- **Wykorzystanie danych do analizy i poprawy procesów wewnętrznych.** Korzystając z dobrych przykładów wskazanych chociażby na przykładzie usługi warto wskazać instytucjom rządowym korzyści płynące z analizy danych i spójnych strategii ich przetwarzania. Kluczowym elementem powinna być dla administracji współpraca w zakresie wykorzystania zasobów danych, a także możliwości wykorzystania posiadanych danych do poprawy świadczonych przez administrację usług.
- **Jasna polityka w sprawie scrapingu.** Niestety wiele zbiorów dostępnych w tym momencie działa na nieujednoliconym interfejsie API, wiele usług dostępnych jest wyłącznie na stronach WWW, a niektórzy usługodawcy wręcz zachęcają do automatycznego pobierania z nich treści (podrozdział **4.2** o scrapingu). Określenie jasnej polityki i wyjaśnienia kwestii prawnych w sprawie wykorzystywania takich narzędzi byłoby dobrą praktyką.

- **Ustalenie standardów lub wzorcowego zestawu udostępnionych danych dla samorządów lokalnych.** Wydaje się zasadne ustalenie kilkupoziomowego standardu otwierania danych, na podstawie przedstawionych w podrozdziale **3.2.3** przykładów polskich i zagranicznych katalogów oferowanych usług.

8.3.2. Rekomendacje dla samorządów

- **Filtrowanie udostępnionych zbiorów danych pod kątem celowości ich wykorzystania, aktualności oraz dublowania danych dostępnych w innych platformach open data.** Większość samorządów umieszcza bardzo wiele opracowań, które można dużo łatwiej pobrać np. z systemów danych GUS. W szczególności dotyczy to zbiorów, które w architekturze GUS są dostępne w dokładniejszej i szerszej formie. Dodatkowo wiele samorządów lokalnych umieszcza dane nie zapewniając ich aktualności. Najlepiej, gdyby aktualizacja taka następowała bezpośrednio na tym samym zbiorze. Samorządy przed publikacją zbioru powinny najpierw odpowiedzieć sobie na pytanie, czy dane publikowane mają charakter unikalny lub mogą być wykorzystane do budowy usług.
- **Publikowanie danych o mniejszym stopniu zagregowania.** Dane powinny być publikowane na najniższym poziomie agregacji, który jest możliwy bez łamania przepisów prawa i bardziej zagregowane opracowania nie powinny być w żadnym przypadku traktowane jako osobny zbiór danych. Idealny byłby model publikacji danych przypominający hurtownie danych z dostępem do poszczególnych miar i wymiarów przez API. **Warto publikować mniej usług, ale o wysokiej jakości, innowacyjności, dostępności i z określonymi potencjalnymi możliwościami wykorzystania.**
- **Łatwiejsze wyszukiwanie katalogu usług dynamicznych API.** Wszystkie usługi udostępnione jako zbiory dynamicznie aktualizowane (API), takie jak aktualne dane o rozkładach jazdy i pozycji autobusów, warstwy geograficzne obiektów w mieście i inne, powinny być umieszczone w osobnych katalogach i wylistowane, wraz z podlinkowaną dokumentacją. Problemem może się stać odnalezienie konkretnej usługi udostępnianej przez samorząd lokalny w przypadku dużej liczby innych zbiorów danych.
- **Unikanie zamykania zbiorów udostępnianych wybranym podmiotom preferencyjnie.** Udostępnianie zbiorów o charakterze publicznym (rozkłady jazdy) jedynie konkretnym twórcom aplikacji na podstawie umów czy indywidualnych ustaleń nie jest dobrą praktyką. Lista dostępnych usług i procedura otrzymania dostępu do nich powinna być jasna i przejrzysta. Unikniemy w ten sposób sytuacji, gdy twórca konkretnego rozwiązania (np. aplikacji informującej o spóźnieniach autobusów) przestanie wspierać aplikację, lub pojawią się błędy. Otwarcie rynku i pozwolenie na rejestrację oraz możliwości testowania API przed przygotowaniem aplikacji, wydają się dużo bardziej zasadne niż tworzenie przetargów na wyprodukowanie konkretnych aplikacji. Tym bardziej, że złożenie oferty na napisanie aplikacji bez wcześniejszego dostępu do API, na którym ma się ona opierać, wydaje się z perspektywy producenta aplikacji sytuacją bardzo niekomfortową. W takim przy-

padku dostawca wcześniej korzystający już ze zbioru danych (np. API transportu publicznego) uzyskuje przewagę nad innymi.

Na podstawie powyższych można zdefiniować krótkie stwierdzenia, które powinny charakteryzować zbiory danych udostępniane przez samorządy lokalne, co można zobaczyć na rysunku **8.8**.

RYSUNEK 8.8. **INFOGRAFIKA: REKOMENDACJE CHARAKTERYZUJĄCE DOBRE ZBIORY OPEN DATA**

Źródło: opracowanie własne.

8.3.3. Rekomendacje dla przedsiębiorców

Cała tematyka raportu, a w szczególności wskazanie dostępnych, użytecznych i aktualnych zbiorów danych jest istocie rekomendacją dla przedsiębiorców. Opisując jednak kwestie nieporuszone bezpośrednio w poprzednich rozdziałach można sformułować kilka rekomendacji na poziomie strategicznym:

- **Współpraca ze studentami i uczelniami w budowaniu narzędzi przetwarzających dane.** Ze względu na charakter danych wymagających często ekstrakcji ze stron HTML, zasadnym wydaje się kontakt z uczelniami i studentami kierunków potrafiących budować takie narzędzia (np. kierunków związanych z przetwarzaniem danych) lub nawet absolwentami techników o profilu informatycznym. Niestety, ze względu na bardzo szybki rozwój liczby udostępnionych zbiorów potrzebny jest pewien zakres umiejętności technicznych w zakresie przetwarzania danych. Dopiero jego posiadanie przez przedsiębiorcę umożliwi stworzenie rozwiązań, które będą mogły służyć budowaniu przewagi konkurencyjnej w postaci ulepszenia istniejących procesów lub pozwoli oferować zupełnie nowe usługi. W związku z powyższym, współpraca z uczelniami w zakresie tworzenia rozwiązań praktycznych jako prace dyplomowe, jak i współpraca na poziomie płatnych praktyk, wydaje się jasną rekomendacją dla pracodawców.
- **Inwestycje w obszary badawcze mające duży potencjał oraz zwiększenie nakładów na R&D.** Obszarami o szczególnym potencjale (nie związanymi bezpośrednio ze specyficznym profilem przedsiębiorstwa) są: smart cities, usługi wykorzystujące open data i machine learning (także deep learning). Możliwe jest także wykorzystanie zbiorów i baz opisanych w raporcie do budowy innowacyjnych usługi i ulepszania istniejących w firmach rozwiązań.

Dodatkowo, opracowano krótki zestaw dobrych praktyk, które mogą być swoistą listą kontrolną dla przedsiębiorców interesujących się wykorzystaniem zasobów danych. Rekomendacje przedstawione zostały na rysunku **8.9**. Rozszerzając kwestie organizacji i wdrażania innowacji (w szczególności nawiązując do punktu 3 z rysunku), poleca się zapoznanie z rozdziałem **5**, w szczególności zaś ze wstępem i metodykami przedstawionymi w sekcji **5.4.2**.

Opisując wskazane na rysunku **8.9** praktyki, aby móc poprawiać lub opracowywać innowacyjne usługi, przedsiębiorcy muszą najpierw być świadomi struktury własnych danych i możliwości ich analizy. Nie każdy przedsiębiorca prowadzi usługi o charakterze technologiczno-informatycznym lub zatrudnia odpowiednią liczbę specjalistów R&D w swojej dziedzinie. Trudno jest też oprzeć całą analitykę na jednym specjaliście – jak wskazaliśmy w sekcji **8.1.1**, do analizy zbiorów danych (abstrahując od ich otwartości) potrzebny jest zestaw wiedzy domenowej i czysto technicznej dotyczący składowania, przetwarzania, analizy i predykcji w związku z analizowanymi zasobami. W związku z powyższym przedsiębiorcy muszą posiadać odpowiedni kapitał ludzki by w ogóle wykorzystać dane istniejące w organizacji. Ważne przy prowadzeniu takich prac jest też **jasne sprecyzowanie oczekiwań** co do podmiotu badań (np. konkretnej firmowej usługi, badania klientów) i celu. Zatrudnienie specjalisty "data science" bez przyjętej strategii dotyczącej zwiększenia innowacyjności nie rozwiąże problemu przedsiębiorstwa, ponieważ każde wdrożenie takiej usługi musi być poparte wcześniejszą analizą problemu (lub szansy) i jasnym celem jego rozwiązania.

RYSUNEK 8.9. **REKOMENDACJE DLA PRZEDSIĘBIORCÓW Z PERSPEKTYWY WYKORZYSTANIA DANYCH W ORGANIZACJI**

Źródło: opracowanie własne.

Pomóc w opracowywaniu innowacyjnych usług może współpraca przedsiębiorców z uczelniami wyższymi: zarówno na poziomie gotowych projektów badawczych, jak i małych konkursów czy organizacji praktyk. Jednak opracowywanie rozwiązań o takiej specyfice wymaga odpowiedniej analizy potrzeb przedsiębiorstwa i akceptacji ryzyka związanego z prowadzeniem działalności badawczej i rozwojowej. Aby akceptować innowacje, firmy muszą zaakceptować długoterminowe skutki prowadzonych badań i analiz, których często nie da się wykazać na poziomie kwartalnych czy nawet rocznych efektów. Bagatelizowanie długoterminowych zysków powoduje niską chęć do wprowadzania innowacji dla wszystkich uczestników tego procesu.

Przedsiębiorcy powinni także wspierać kulturą organizacyjną opracowywania nowych rozwiązań, nawet gdy są to małe projekty oparte na danych. Korzystając z wzorców firm technologicznych takich, jak Google czy Facebook, przedsiębiorstwa mogą zachęcać pracowników do

opracowywania ciekawych rozwiązań, np. podczas jednego dnia ich tygodniowej pracy. Rozwiązania te mogą dotyczyć nie tylko rozwiązań informatycznych, ale także innych dziedzin. Jest to rodzaj kultury organizacyjnej wspierającej innowacje w przedsiębiorstwie, a nie skupionej na maksymalizacji godzin pracy pracownika. Pewna swoboda i stabilność możliwości prowadzenia prac badawczych i rozwojowych sprzyja opracowywaniu innowacyjnych rozwiązań. Zaprezentowane rekomendacje oczywiście nie rozwiążą wszystkich problemów, ale mogą pozwolić przedsiębiorcom na opracowanie realnych oczekiwań i strategii względem wykorzystania otwartych i wewnętrznych zasobów danych.

8.3.4. Istotne obszary przetwarzania danych

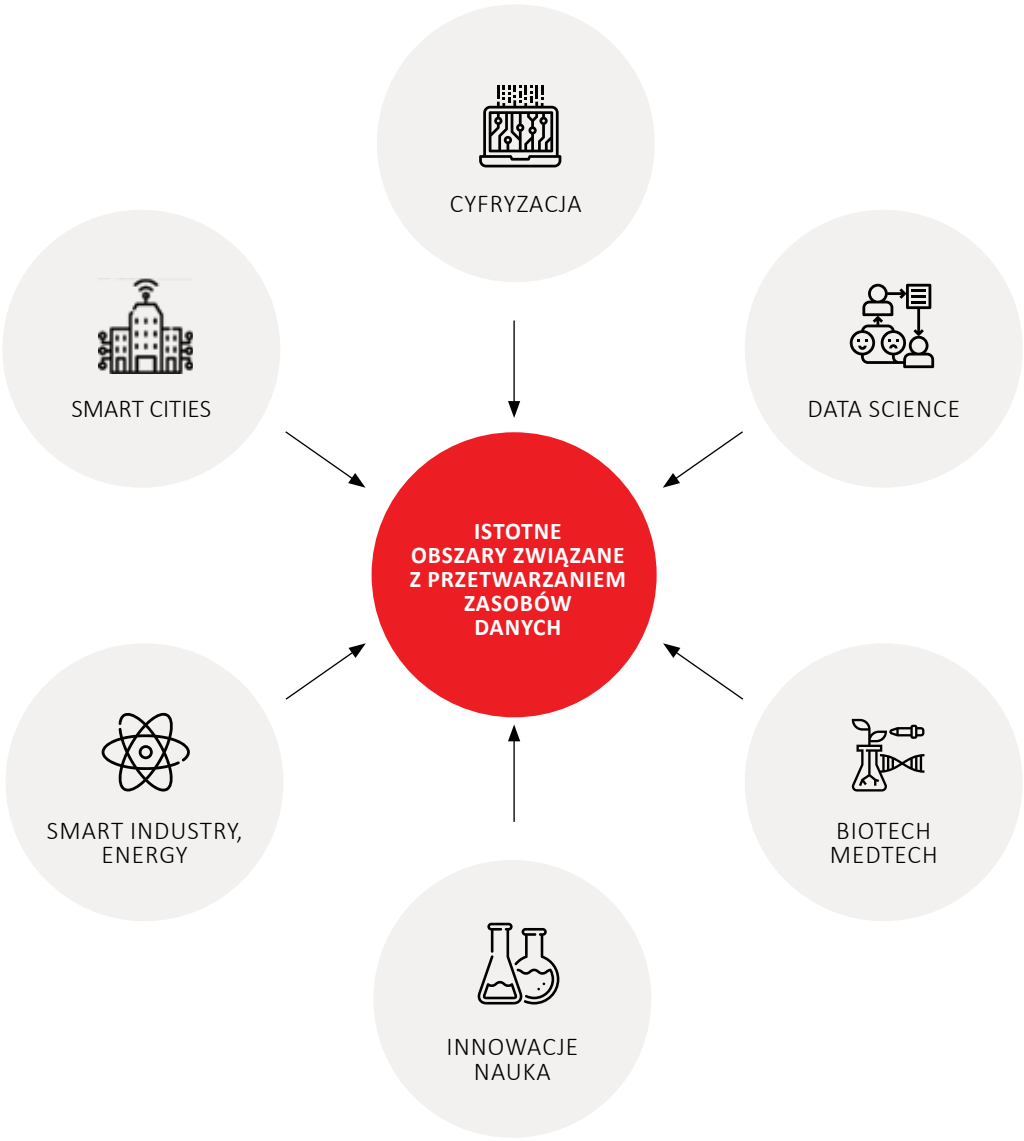
Tworzenie wartości z wykorzystaniem zasobów danych **wymaga współpracy** pomiędzy administracją, przedsiębiorcami, jak i sektorem naukowym i edukacyjnym. Bazując na zidentyfikowanych w raporcie zasobach można wyróżnić kilka najistotniejszych trendów wartych uwagi z perspektywy wykorzystania danych. Trendy te zostały zidentyfikowane na podstawie analizy przykładów usług i badań naukowych, są to również obszary technologiczne charakteryzujące się dużym poziomem innowacyjności. Przykłady te, zaprezentowane na rysunku **8.10**, wyznaczają potencjalne rekomendacje dla wszystkich zainteresowanych:

- **Smart Cities** – w przypadku inteligentnych rozwiązań miejskich mowa tutaj zarówno o budowie i dostarczaniu sensorów fizycznych, poprzez opracowywanie rozwiązań analitycznych i wdrażanie ich w różnej skali (na poziomie gospodarstwa domowego, budynku, jak i całej infrastruktury miejskiej). Zastosowane metody mogą także obejmować analitykę wykorzystującą otwarte zbiory danych, takie jak OpenStreetMap. Możliwe jest także wyróżnienie obszarów tematycznych, tj.: transport, ekologia i ochrona środowiska, IoT (Internet Rzeczy) i inteligentne urządzenia zarządzania domem, wykorzystanie różnych źródeł danych (np. social media, sensory natężenia ruchu i hałasu) do oceny transportu, zagospodarowania przestrzennego i jakości życia obywateli. Rozwiązania z zakresu Smart Cities powinny być raczej skupione na szczeblu lokalnym i badaniach pilotażowych (stąd rozróżnienie z tematyką poruszoną poniżej).
- **Smart Industry (Industry 4.0) i Smart Energy** – unowocześnianie kluczowych sektorów polskiej gospodarki za pomocą nowych technologii. To inwestycja nie tylko w nowe rozwiązania, ale także ich krytyczna analiza pod kątem bezpieczeństwa i opracowywanie rozwiązań pozwalających na efektywniejsze zarządzanie zasobami. Wspieranie współpracy pomiędzy przedsiębiorstwami, administracją i nauką (zarówno w perspektywie projektów, jak i aktywizacji studentów). Rozwiązania z zakresu Smart Industry i Energy powinny być raczej koordynowane na szczeblu krajowym (stąd rozróżnienie z tematyką wspomnianą powyżej).
- **BioTech i MedTech** – w szczególności inwestycje w kluczowe kompetencje i naukowców skupionych w tych obszarach, premiowanie badań naukowych i patentów, zabezpieczanie

zatrudnienia młodym specjalistom. W przypadku MedTech także premiowanie badań interdyscyplinarnych: z wykorzystaniem nowych źródeł danych (jak smartwatche, sensory), dotyczącymi profilaktyki zdrowotnej i tematów interesujących dla przedsiębiorców (np. badanie poziomu stresu w miejscu pracy).

- **Cyfryzacja** – rozumiana zgodnie z ideą open data, ale skierowana też wewnętrznie dla poprawy jakości usług administracji. Działania mające na celu poszerzanie współpracy i koordynację działań pomiędzy strukturami administracyjnymi zarządzającymi usługami udostępniającymi zbiory danych. Skupienie się na celowości i istotnych zasobach danych, którymi państwo dysponuje oraz analiza zasobów udostępnianych cyfrowo. Wspieranie rozwiązań otwierających szczegółowe, dynamiczne zasoby danych i inwestycje w aparaturę (także na szczeblach samorządowych) pozwalającą na udostępnianie takich usług.
- **Data Science** – inwestycja w obszary badawcze i dydaktyczne związane z analizą danych. Krytyczne są obszary pozwalające na **analizę i automatyczne podejmowanie decyzji na podstawie różnorodnych danych o dużym wolumenie**, co wiąże się z tematami takimi, jak: big data, machine learning (uczenie maszynowe) z włączeniem metod przetwarzania tekstu, artificial intelligence (sztuczna inteligencja), deep learning (metody głębokiego uczenia). Dodatkowo warto podejmować działania będące próbą przeciwdziałania drenażowi wiedzy z polskiej gospodarki, w celu zatrzymania specjalistów zajmujących się istotnymi obszarami badawczymi. Potencjalne rekomendacje mogą obejmować więcej programów naukowych dla młodych naukowców i wspieranie inicjatyw takich jak: praktyki, staże i doktoraty wdrożeniowe w obszarach związanych z analizą danych oraz R&D. Rola administracji jako dostawcy dużych i niezagregowanych danych do analiz na potrzeby projektów dla podmiotów zewnętrznych wydaje się także interesującym kierunkiem dla rozwoju tego obszaru.
- **Nauka i innowacje** – obszar obejmujący celowe działania w celu analizy i wspierania osiągnięć z zakresu nauki i wdrażania innowacyjnych rozwiązań, a także wspierania działań istotnych dla modelu innowacyjnej gospodarki. Przykładem takiego działania może być analiza aktualnego stanu polskiej nauki i techniki (podana także jako przykład w raporcie), celem identyfikacji kluczowych kompetencji i obszarów badań. Pomocne w popularyzacji innowacji może być przygotowywanie konkursów, zgodnie z rekomendacjami z raportu dotyczącymi open innovation. Kluczowe jest zacieśnienie współpracy pomiędzy polskimi przedsiębiorstwami a uczelniami i studentami oraz promowanie współpracy z instytucjami i administracją rządową.

RYSUNEK 8.10. **INTERESUJĄCE OBSZARY Z PERSPEKTYWY ISTOTNOŚCI ZASOBÓW I ANALIZY DANYCH**



Źródło: opracowanie własne.

8.3.5. Podsumowanie

W niniejszym raporcie dokonano przeglądu dostępnych zbiorów danych oraz oceniono ich przydatność. Wskazano różnorodność typów i źródeł pochodzenia danych, wraz z możliwościami ich wykorzystania do budowania usług i aplikacji. Wskazano jednocześnie na najciekawsze i najbardziej wartościowe źródła i zasoby. Wskazano również sposoby pozyskiwania danych, opisując różnorodne standardy ich przechowywania i narzędzia oraz metody mogące służyć ich wstępnej obróbce. Ze względu na wolumen danych wskazano istotność metod takich jak uczenie maszynowe i metody sztucznej inteligencji. Zaprezentowano również analizę przypadków ilustrującą różne scenariusze wykorzystania i przetwarzania danych. Przedstawione w raporcie treści pozwalają odpowiedzieć na najistotniejsze pytania związane z kierunkami rozwoju polskich przedsiębiorstw najintensywniej pracujących z danymi. Tak szerokie opracowanie może być dobrym punktem wyjścia dla opracowania wszelakiego rodzaju strategii szczegółowych w zakresie: otwartych danych, data science, kształtowania strategii edukacji i kluczowych kompetencji w gospodarce cyfrowej czy road mapy rozwoju sztucznej inteligencji w Polsce.

Bibliografia

[1] Witold Abramowicz. *Filtrowanie informacji*. Wydawnictwo Akademii Ekonomicznej, 2008.

[2] Abel Gezevich Aganbegjan. "Dostizhenie vysshego urovnja prodolzhitelnosti zhizni v Rossii". W: *Rossijskoe predprinimatelstvo* 2 (2012).

[3] Albin Ahmeti i inni. "Updating Wikipedia via DBpedia Mappings and SPARQL". W: *The Semantic Web*. Redagował Eva Blomqvist i inni. Cham: Springer International Publishing, 2017, strony 485–501. : -319-58068-5.

[4] Alexa. *wikia.com Traffic Statistics*. 2018.: https://www.alexa.com/ siteinfo/wikia.com (dostęp 08.10.2018).

[5] Mustafa Aljumaili, Ramin Karim i Phillip Tretten. "Quality of Streaming Data in Condition Monitoring Using ISO 8000". W: *Current Trends in Reliability, Availability, Maintainability and Safety*. Redagował Uday Kumar i inni. Cham: Springer International Publishing, 2016, strony 703–715. :-319-23597-4.

[6] Milad Alshomary i inni. "Wikipedia Text Reuse: Within and Without". W: *arXiv preprint arXiv:1812.09221* (2018).

[7] Benjamin M Althouse i inni. "Differences in impact factor across fields and over time-".W: *Journal of theAmerican Society for Information Science and technology* 60.1 (2009), strony 27–34.

[8] Imaduddin Amin i inni. "Inferring commuting statistics in greater Jakarta from social media locational information from mobile devices". W: (2017).

[9] Maik Anderka. "Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia". PhD. Bauhaus-Universitaet Weimar Germany, 2013.: http:// www.uni-weimar.de/medien/webis/ publications/papers/anderka%5C_2013.pdf.

[10] Alessio Palmero Aprosio, Claudio Giuliano i Alberto Lavelli. "Towards an automatic creation of localized versions of DBpedia". W: *International Semantic Web Conference*. Springer. 2013, strony 494–509.

[11] Mathias Astell i inni. *Practical challenges for researchers in data sharing -Springer Nature survey data (anonymised)*. 2018. : https://figshare.com/articles/Practical_challenges_ for _ researchers _ in _ data _ sharing_ -_Springer_Nature_survey_data_anonymised_/5971387 (dostęp06.02.2019).

[12] AT&T. *A definitive list of smart cities with open APIs*. 2017. : https:// developer.att.com/ blog/smart-cities-with-open-apis (dostęp 08.01.2019).

[13] Judie Attard, Fabrizio Orlandi i Sören Auer. "Value creation on open government data". W: *System Sciences (HICSS), 2016 49th Hawaii International Conference on*. IEEE. 2016, strony 2605–2614.

[14] Judie Attard i inni. "A systematic review of open government data initiatives". W: *Government Information Quarterly* 32.4 (2015), strony 399–418.

[15] DE Avison i G Fitzgerald. "Information Systems Development: Methodologies, Techniques and Tools". W: (1995).

[16] Carlo Batini i Monica Scannapieco. "Data Quality Dimensions". W: *Data and Information Quality: Dimensions, Principles and Techniques*. Cham: Springer International Publishing, 2016, strony 21–51. : 978-3-319-24106-7. : 10.1007/978-3-319-24106-7_2.

[17] Christian Becker i Christian Bizer. "DBpedia Mobile:A Location-Enabled Linked Data Browser." W: *Ldow* 369 (2008), strona 2008.

[18] P Benson. "NATO codification system as the foundation for ISO 8000, the international standard for data quality". W: *Oil IT Journal* (2008).

[19] Jorn Berends, Wendy Carrara i Cosmina Radu. *European Data Portal Analytical Report 9: European Data Portal*. Sprawozdanie techniczne. 2017. : https://www.europeandataportal.eu/sites/default/ files/analytical_report_n9_economic_benefits_of_open_data.pdf.

[20] Jorn Berends, Wendy Carrara i Heleen Voller. *Analytical Report 6: Open Data in Cities 2*. 2016. : https://www.europeandataportal.eu/sites/ default/files/edp_analytical_report_n6_-_open_data_ in_cities_2_-_final-clean.pdf (dostęp 08.01.2019).

[21] JanekBevendorffiinni."ElasticChatNoir:SearchEnginefortheClueWebandthe Common Crawl". W: *European Conference on Information Retrieval*. Springer. 2018, strony 820–824.

[22] Christian Bizer i inni. "DBpedia-A crystallization point for the Web of Data". W: *Web Semantics: science, services and agents on the world wide web* 7.3 (2009), strony 154–165.

[23] Christian Bizer i inni. "Deployment of rdfa, microdata, and microformats on the web–a quantitative analysis". W: *International Semantic Web Conference*. Springer. 2013, strony 17–32.

[24] Joshua E Blumenstock. *Automatically Assessing the Quality of Wikipedia Articles*. Sprawozdanie techniczne. 2008.: 101080/17439880802324251.: http://repositories.cdlib.org/cgi/viewcontent.cgi? article=1021%5C&context=ischool.

[25] Joshua E Blumenstock. "Size matters: word count as a measure of quality on wikipedia". W: *WWW*. 2008, strony 1095–1096. : 9781605580852.: 10. 1145 / 1367497. 1367673. : http://portal.acm.org/citation.cfm?id=1367673.

[26] Łukasz Borowiecki i Piotr Mieczkowski. *Map of the Polish AI*. Warszawa: Digital Poland Foundation, 2019. : 978-83-951530-1-3. : https://www.digitalpoland.org/assets/reports/map-ofthe-polish-ai---2019-edition-i.pdf (dostęp 04.02.2019).

[27] Samir K Brahmachari. *A Joint Front*. 2013. : https://www.businesstoday.in/magazine/the-buzz/ajoint-front/story/304076.html (dostęp 06.02.2019).

[28] Thierry Buecheler i inni. "Crowdsourcing, Open Innovation and Collective Intelligence in the Scientific Method-A Research Agenda and Operational Framework." W: *ALIFE*. 2010, strony 679–686.

[29] Center for Open Data Enterprise. *Open Data Impact Map*. Sprawozdanie techniczne May. 2016, strony 1–28.

[30] H. Chesbrough. *Open Innovation: The New Imperative for Creating and Profiting, from Technology*. Boston: Harvard Business School Press, 2003.

[31] C Clare i P Loucopoulos. "Business Information Systems". W: (1987).

[32] Commecica.com. *The prevalence of Web advertising*. 2018. : https://commecica.com/2018/06/27/web-ad-prevalence/ (dostęp 27.01.2019).

[33] Commission of the European Communities. *eEurope 2002: Quality criteria for health related websites*. 2002. : 10.2196/jmir.4.3.e15.

[34] Common Crawl. *Common Crawl Index Server*. 2019. : http://index. commoncrawl.org/ (dostęp 17.01.2019).

[35] Riccardo Conti i inni. "Maturity assessment of Wikipedia medical articles". W: *Computer-Based Medical Systems (CBMS), 2014 IEEE 27th International Symposium on*. IEEE. 2014, strony 281–286.

[36] Holly Crawford. "Encyclopedias". W: *Reference and information services: An introduction* (2001), strony 433–459.

[37] Edward Curry. "The big data value chain: definitions, concepts, and theoretical approaches". W: *New horizons for a data-driven economy*. Springer, Cham, 2016, strony 29–37.

[38] Ministerstwo Cyfryzacji. *Nowa strona danych – Hackathon Ministerstwa Cyfryzacji*. 2018. : https://dane.gov.pl/article/1217 (dostęp 08.01.2019).

[39] Ministerstwo Cyfryzacji. *Pierwszy rządowy hackathon: przełom w podejściu do danych publicznych*. 2018. : https://www.gov.pl/web/cyfryzacja/pierwszy-rzadowyhackathon-przelom-w-podejsciu-do-danych-publicznych (dostęp 08.01.2019).

[40] Daniel Hasan Dalip i inni. "Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia". W: *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*. 2009, strony 295–304. : 9781605583228. : 10.1145/1555400.1555449. : http://portal.acm.org/citation.cfm?id=1555400.1555449.

[41] Daniel Hasan Dalip i inni. "Automatic Assessment of Document Quality in Web Collaborative Digital Libraries". W: *Journal of Data and Information Quality* 2.3 (2011), strony 1–30. : 19361955. : 10.1145/2063504.2063507.

[42] Daniel H Dalip i inni. "A general multiview framework for assessing the quality of collaboratively created content on web 2.0". W: *Journal of the Association for Information Science and Technology* 68.2 (2017), strony 286–308.

[43] Quang-Vinh Dang i Claudia-Lavinia Ignat. "Quality assessment of Wikipedia articles without feature engineering". W: *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*. 2016, strony 27–30.

[44] J. Deng i inni. "ImageNet: A Large-Scale Hierarchical Image Database". W: *CVPR09*. 2009.

[45] Zhonghua Deng i Ling Luo. "An exploratory discuss of new ways for competitive intelligence on Web2. 0". W: *Integration and Innovation Orient to E-Society Volume 2*. Springer, 2007, strony 597–604.

[46] Rex W Douglass i inni. "High resolution population estimates from telecommunications data". W: *EPJ Data Science* 4.1 (2015), strona 4.

[47] EFQM i Fundacja Rozwoju Uniwersytetu Ekonomicznego we Wrocławiu. *Model doskonałoci EFQM*. 2013.

[48] Kemele M Endris i inni. "Dataset reuse: an analysis of references in community discussions, publications and data". W: *Proceedings of the Knowledge Capture Conference*. ACM. 2017, strona 5.

[49] English Wikipedia. *API sandbox*. 2019. : https://en.wikipedia.org/wiki/Special:ApiSandbox (dostęp 03.02.2019).

[50] M.J. Eppler. *Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes*. Springer Berlin Heidelberg, 2013. : 9783540247821. : https://books.google.pl/books?id=ChoHCAAAQBAJ.

[51] Fandom. *Explore*. 2018. : http://fandom.wikia.com/explore (dostęp 08.10.2018).

[52] David Ferrucci i inni. "Building Watson: An overview of the DeepQA project". W: *AI magazine* 31.3 (2010), strony 59–79.

[53] Tiziano Flati i inni. "MultiWiBi:The multilingualWikipedia bitaxonomy project". W: *Artificial Intelligence* 241 (2016), strony 66–102.

[54] Sławomir Folwarski i Mateusz Ossowski. "Być w 15%". W: *Nowa twarz business intelligence* (2012), strony 64–68.

[55] Fundacja ePaństwo. *Sejmometr API*. 2018. : http://kdpforum. s3. amazonaws. com / 5643ba1b392450aa7882d0f2c9502188a049c9cf5f. pdf (dostęp 06.02.2019).

[56] Frank Galland. *Dictionary of Computing: Data Communications, Hardware and Software Basics, Digital Electronics*. John Wiley & Sons Inc, 1982. : 047110468X.

[57] geoportal.gov.pl. *Dane ewidencyjne*. 2018. : https://www.geoportal. gov.pl/dane/dane--ewidencyjne (dostęp 01.02.2019).

[58] geoportal.gov.pl. *Usługi przeglądania WMS*. 2018. : https://www.geoportal.gov.pl/uslugi/usluga-przegladania-wms (dostęp 01.02.2019).

[59] Jim Giles. *Internet encyclopaedias go head to head*. 2005.

[60] GIS-NET. *Lista Serwisów WMS*. 2018. : http://www.gis-net.pl/index.php?option=com_content& view=article&id=61&Itemid=99 (dostęp 01.02.2019).

[61] Wolfgang Glänzel i András Schubert. "A new classification scheme of science fields and subfields designed for scientometric evaluation purposes". W: *Scientometrics* 56.3 (2003), strony 357–367.

[62] Urząd Statystyczny Główny. *Hackathon GUS: Show me data*. 2018. :https://hackathon.stat.gov.pl/ (dostęp 01.02.2019).

[63] Mariusz Grabowski i Agnieszka Zając. "Dane, informacja, wiedza–próba definicji". W: *Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie* 798 (2009), strony 99–116.

[64] Emily Grantner. "ISO 8000: a standard for data quality". W: *Logistics Spectrum* 41.4 (2007).

[65] Independent Expert Advisory Group. *A World that Counts. Mobilising the data revolution for sustainable development*. Sprawozdanie techniczne. United Nations, 2014, strona 28. : http://www.undatarevolution.org.

[66] Richard D. Hackathorn. *Web Farming for the Data Warehouse (The Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann, 2001. : 05037. : https://www.amazon.com/Farming-Warehouse-KaufmannManagement-Systems/dp/1558605037?SubscriptionId= AKIAIOBINVZYXZQZ2U3A&tag=chimbori05--20&linkCode=xm2& camp=2025&creative=165953&creativeASIN=1558605037.

[67] Aaron Halfaker. "Interpolating quality dynamics in wikipedia and demonstrating the keilana effect". W: *Proceedings of the 13th International Symposium on Open Collaboration*. ACM. 2017, strona 19.

[68] Bin He i inni. "Accessing the deep web". W: *Communications of the ACM* 50.5 (2007), strony 94–101.

[69] Bernd Heinrich, Marcus Kaiser i Mathias Klier. "How to measure data quality? A metric--based approach". W: (2007).

[70] Enrique Herrera-Viedma i inni. "Evaluating the information quality of web sites: A methodology based on fuzzy computing with words". W: *Journal of the American Society for Information Science and Technology* 57.4 (2006), strony 538–549.

[71] Sven Hertling i Heiko Paulheim. "Dbkwik:A consolidated knowledge graph from thousands of wikis". W: *2018 IEEE International Conference on Big Knowledge (ICBK)*. IEEE. 2018, strony 17–24.

[72] James O. Hicks. *Management Information Systems: A User Perspective*. West Group, 1993. : 0314933670.

[73] Anders Hjalmarsson, Gustaf Juell-Skielse i Paul Johannesson. "Open Digital Innovation Contest". W: *Open Digital Innovation: A Contest Driven Approach*. Cham: Springer International Publishing, 2017, strony 11–21. : 978-3-319-56339-8. : 10.1007/978-3-319-56339-8_3. : https://doi.org/10.1007/978-3-319-56339-8_3.

[74] Anders Hjalmarsson, Gustaf Juell-Skielse i Paul Johannesson. *Open Digital Innovation: A Contest Driven Approach*. Cham: Springer International Publishing, 2017. : 978-3-319-56339-8. : 10. 1007 / 978 – 3 – 319 – 56339 – 8 _ 3. : https://doi.org/10.1007/978-3-319-56339-8_3.

[75] Anders Hjalmarsson, Gustaf Juell-Skielse i Paul Johannesson. "Open Innovation". W: *Open Digital Innovation: A Contest Driven Approach*. Cham: Springer International Publishing, 2017, strony 5–9. : 978-3-319-56339-8. : 10. 1007 / 978 – 3 – 319 – 56339 – 8 _ 2. : https://doi.org/10.1007/978-3-319-56339-8_2.

[76] S Holwell i P Checkland. *Information, Systems and Information Systems: Making sense of the field*. 2002.

[77] Meiqun Hu i inni. "Measuring article quality in wikipedia". W: *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management -CIKM '07*. 2007, strony 243–252. : 9781595938039. : 10.1145/1321440.1321476. : http://portal.acm.org/ citation. cfm ? id = 1321476 $ %5Cbackslash $ nhttp ://portal.acm.org/citation.cfm?doid=1321440.1321476.

[78] ISO/TS. *Technical specification ISO/TS 8000-150:2011(E)—data quality—Part 1: overview*. Geneva, Switzerland, 2011.

[79] Marijn Janssen, Yannis Charalabidis i Anneke Zuiderwijk. "Benefits, adoption barriers and myths of open data and open government". W: *Information systems management* 29.4 (2012), strony 258–268.

[80] Marijn Janssen, Ricardo Matheus i Anneke Zuiderwijk. "Big and open linked data (BOLD) to create smart cities and citizens: Insights from smart energy and mobility cases". W: *International Conference on Electronic Government*. Springer. 2015, strony 79–90.

[81] Marijn Janssen i inni. "Driving public sector innovation using big and open linked data (BOLD)". W: *Information Systems Frontiers* 19.2 (2017), strony 189–195.

[82] Yu-ying Jiao i Jing Yuan. "Research on the Collective Knowledge Sharing and Innovation Service Based on WIKI [J]". W: *Information Science* 5 (2008), strona 003.

[83] Michael Johnston. *Wikipedia Revenue Analysis: How a Wiki Could Make $2.3B a Year*. 2013. : https://monetizepros.com/features/analysis-howwikipedia-could-make-2-8-billion-in-annual-revenue/ (dostęp 08.10.2018).

[84] Chichang Jou. "Schema Extraction for Deep Web Query Interfaces Using Heuristics Rules". W: *Information Systems Frontiers* (2018), strony 1–12.

[85] Yasin Kabalci. "A survey on smart metering and smart grid communication". W: *Renewable and Sustainable Energy Reviews* 57 (2016), strony 302–318.

[86] Piotr Kałużny i Piotr Stolarski. "Biometria behawioralna i „tradycyjna" w mobilnych usługach bankowych – stan oraz przyszłe możliwości zastosowania". pl. W: *Bezpieczny Bank* 1 (2019), strony 139–161. : 10.26354/bb.7.1.74.2019.

[87] Soyoung Kim i Leslie Stoel. "Apparel retailers: website quality dimensions and satisfaction". W: *Journal of Retailing and Consumer Services* 11.2 (2004), strony 109–117.

[88] Gary King. "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing". W: *Sociological Methods and Research* 36 (2007), strony 173–199.

[89] Alice Valle Knight i David J. Silk. *Managing Information: Information Systems for Today's General Manager (Henley Management Series)*. McGraw-Hill, 1990. : 0077070860.

[90] Shinji Kobayashi, Thomas B Kane i Chris Paton. "The privacy and security implications of open data in healthcare". W: *Yearbook of medical informatics* (2018).

[91] Komisja Europejska. *Indeks gospodarki cyfrowej i społeczeństwa cyfrowego (DESI). Sprawozdanie krajowe na 2019 r. Polska*. 2018. : https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60000.

[92] Europejska Komisja. *Proposal for a revision of the Public Sector Information (PSI) Directive*. 2018. : https://ec.europa.eu/digitalsingle-market/en/proposal-revision-public--sectorinformation–psi–directive (dostęp 08.01.2019).

[93] Günter Krampen i inni. "On the validity of citation counting in science evaluation: Content analyses of references and citations in psychological publications". W: *Scientometrics* 71.2 (2007), strony 191–202.

[94] Kenneth C. Laudon i Jane P. Laudon. *Business Information Systems: A ProblemSolving Approach*. Dryden Pr, 1991. : 0030988179.

[95] Jens Lehmann i inni. "DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia". W: *Semantic Web* 6.2 (2015), strony 167–195.

[96] Włodzimierz Lewoniewski. "Enrichment of Information in Multilingual Wikipedia Based on Quality Analysis". W: *International Conference on Business Information Systems*. Springer. 2017, strony 216–227.

[97] Włodzimierz Lewoniewski. "Measures for Quality Assessment of Articles and Infoboxes in Multilingual Wikipedia". W: *International Conference on Business Information Systems*. Springer. 2019, strony 619–633.

[98] Włodzimierz Lewoniewski i Krzysztof Węcel. "Relative Quality Assessment of Wikipedia Articles in Different Languages Using Synthetic Measure". W: *Business Information Systems Workshops: BIS 2017 International Workshops, Poznań, Poland, June 28-30, 2017, Revised Papers*. Redagował Witold Abramowicz. Cham: Springer International Publishing, 2017, strony 282–292. : 978-3-319-69023-0. : 10.1007/978-3-319-69023-0_24.

[99] Włodzimierz Lewoniewski, Krzysztof Węcel i Witold Abramowicz. "Analiza porównawcza modeli jakości informacji w narodowych wersjach Wikipedii". W: *Systemy Wspomagania Organizacji SWO 2015*. Redagował Teresa Porębska-Miąc. Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, 2015, strony 133–154.

[100] Włodzimierz Lewoniewski, Krzysztof Węcel i Witold Abramowicz. "Quality and Importance of Wikipedia Articles in Different Languages". W: *Information and Software Technologies: 22nd International Conference, ICIST 2016, Druskininkai, Lithuania, October 13-15, 2016, Proceedings*. Cham: Springer International Publishing, 2016, strony 613–624. : 978-3-319-46254-7. : 10.1007/978-3-319-46254-7_50.

[101] Włodzimierz Lewoniewski, Krzysztof Węcel i Witold Abramowicz. "Analysis of References Across Wikipedia Languages". W: *Information and Software Technologies: 23 rd International Conference, ICIST 2017, Druskininkai, Lithuania, October 12–14, 2017, Proceedings*. Redagował Robertas Damasevicius i Vilma Mikayte. Cham: Springer International Publishing, 2017, strony 561–573. : 978-3-319-67642-5. : 10. 1007 / 978 – 3 – 319 – 67642 – 5 _ 47. : https://doi.org/10.1007/978-3-319-67642-5_47.

[102] Włodzimierz Lewoniewski, Krzysztof Węcel i Witold Abramowicz. "Relative Quality and Popularity Evaluation of Multilingual Wikipedia Articles". W: *Informatics* 4.4 (2017). : 2227-9709. :10.3390/informatics4040043.

[103] Włodzimierz Lewoniewski i inni. "Kompletność danych o produktach w infoboksach różnych wersji językowych Wikipedii". W: *Studia Oeconomica Posnaniensia* 6.9 (2018), strony 39–58. : 10.18559/SOEP.2018.9.3.

[104] Elisabeth Lex i inni. "Measuring the quality of web content using factual information". W: *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality - WebQuality '12* (2012), strona 7. : 10. 1145 / 2184305. 2184308. : http://dl.acm.org/citation.cfm?id=2184305.2184308.

[105] Han Li i inni. "Analyzing housing prices in Shanghai with open data: Amenity, accessibility and urban structure". W: *Cities* 91 (2019), strony 165–179.

[106] Libelium. *50 Sensor Applications for a Smarter World*. 2018. : http://www.libelium.com/resources/top_50_iot_sensor_ applications_ranking/ (dostęp 08.01.2019).

[107] Andrew Lih. "Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource". W: *5th International Symposium on Online Journalism* (2004), strona 31.

[108] Juho Lindman, Tomi Kinnari i Matti Rossi. "Industrial open data: Case studies of early open data entrepreneurs". W: *2014 47th Hawaii International Conference on System Sciences (HICSS)*. IEEE. 2014, strony 739–748.

[109] Nedim Lipka i Benno Stein. "Identifying Featured Articles in Wikipedia: Writing Style Matters". W: *Proceedings of the 19th International Conference on World Wide Web (2010)* (2010), strony 1147–1148. : 10.1145/1772690.1772847.

[110] Xingjian Liu i inni. "Understanding urban China with open data". W: *Cities* 47 (2015), strony 53–61.

[111] R.N Maddison. *Information Systems Development for Managers*. Paradigm, London, 1989.

[112] Jayant Madhavan i inni. "Google's deep web crawl". W: *Proceedings of the VLDB Endowment* 1.2 (2008), strony 1241–1252.

[113] Stuart Madnick i inni. "Overview and Framework for Data and Information Quality Research". W: *ACM Journal of Data and Information Quality* 1.1 (2009), strony 1–22. : 1936-1955. : 10.1145/1515693.1516680.http.

[114] Gustavo Magalhaes i Catarina Roseira. "Open government data and the private sector: an empirical view on business models and value creation". W: *Government Information Quarterly* (2017).

[115] David Matthews. *Do academic social networks share academics' interests?* 2016. : https://web.archive.org/web/20160417100025/https: //www.timeshighereducation.com/features/do-academicsocial – networks – share – academics – interests (dostęp 06.02.2019).

[116] Meta S. Brown, Forbes. *City Governments Making Public Data Easier To Get: 90 Municipal Open Data Portals*. 2018. : https ://www.forbes.com/sites/metabrown/2018/04/29/citygovernments-making-public-data-easier-to-get-90municipal – open – data – portals / #396139735a0d (dostęp 08.01.2019).

[117] Robert Meusel i inni. "Graph structure in the web—revisited: a trick of the heavy tail". W: *Proceedings of the 23 rd international conference on World Wide Web*. ACM. 2014, strony 427–432.

[118] Ministerstwo Cyfryzacji. *PROGRAM OTWIERANIA DANYCH PUBLICZNYCH – Załącznik do uchwały nr 107/2016 Rady Ministrów z dnia 20 września 2016 r.* 2016. : https://dane.gov.pl/media/ckeditor/2018/11/22/ otwieranie-danych-podrecznik-dobrych-praktyk.pdf (dostęp 01.02.2019).

[119] Ministerstwo Cyfryzacji. *Otwieranie Danych – Protokół z otwarcia projektu*. 2017. : https://mc.bip.gov.pl/fobjects/download/207914/201701-19-protokol-z-prezentacji-publicznej-projektuotwarte-dane-pdf.html (dostęp 01.02.2019).

[120] Ministerstwo Cyfryzacji. *Otwieranie Danych – Podręcznik dobrych praktyk*. 2018. : https://dane.gov.pl/media/ckeditor/2018/11/22/ otwieranie-danych-podrecznik--dobrych-praktyk.pdf (dostęp 01.02.2019).

[121] Ministerstwo Cyfryzacji. *Standardu interfejsu programistycznego aplikacji.* 2018. : https://www.gov.pl/web/cyfryzacja/standardinterfejsu – programistycznego – aplikacji – api (dostęp 01.02.2019).

[122] Blaise Misztal. "Strategiczne myślenie w cyberbezpieczeństwie". W: *CyberDefence24* (2016). : https://www.cyberdefence24.pl/strategiczne-mysleniew-cyberbezpieczenstwie (dostęp 27.09.2019).

[123] RP Mohanty, D Seth i S Mukadam. "Quality dimensions of e-commerce and their implications". W: *Total Quality Management & Business Excellence* 18.3 (2007), strony 219–247.

[124] Andrea Moro i Roberto Navigli. "SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking". W: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado: Association for Computational Linguistics, 2015, strony 288–297. : http://www.aclweb.org/anthology/S15-2049.

[125] MPiT. *Mapa drogowa rozwoju sztucznej inteligencji w Polsce*. MPiT. 2019. : https://www.gov.pl/web/przedsiebiorczosc technologia/konferencja – mapa – drogowa – sztucznej inteligencji-w-polsce (dostęp 04.02.2019).

[126] Domenico Natale. "Complexity and data quality". W: *Poster e Atti Conferenza*. 2011, strony 13–16.

[127] Roberto Navigli i Simone Paolo Ponzetto. "BabelNet:The automatic construction, evaluation and application of a wide-coverage multilingual semantic network".W: *Artificial Intelligence* 193 (2012), strony 217–250.

[128] Roberto Navigli i Simone Paolo Ponzetto. "Joining Forces Pays Off: Multilingual Joint Word Sense Disambiguation". W: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju, Korea, 2012, strony 1399–1410.

[129] Roberto Navigli i Simone Paolo Ponzetto. "Multilingual WSD with Just a Few Lines of Code: the BabelNet API". W: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*. Jeju, Korea, 2012.

[130] Pascal Neis, Dennis Zielstra i Alexander Zipf. "The street network evolution of crowdsourcedmaps: OpenStreetMapin Germany 2007–2011".W:*FutureInternet* 4.1 (2011), strony 1–21.

[131] Sebastian Neumaier, Jürgen Umbrich i Axel Polleres. "Automated quality assessment of metadata across open data portals". W: *Journal of Data and Information Quality (JDIQ)* 8.1 (2016), strona 2.

[132] Oliver O'Brien i James Cheshire. "Interactive mapping for large, open demographic data sets using familiar geographical features". W: *Journal of Maps* 12.4 (2016), strony 676–683.

[133] Randal S. Olson i inni. "PMLB: a large benchmark suite for machine learning evaluation and comparison". W: *BioData Mining* 10.1 (2017), strona 36. : 1756-0381. : 10.1186/s13040-017-0154-4. : https://doi. org/10.1186/s13040-017-0154-4.

[134] OpenStreetMap Foundation. *Licence*. 2019. : https://wiki.osmfoundation.org/wiki/Licence (dostęp 26.01.2019).

[135] OpenStreetMap Foundation. *Using OpenStreetMap*. 2019. : https://wiki.openstreetmap.org/wiki/Using_OpenStreetMap (dostęp 26.01.2019).

[136] Albert Opher i inni. "The Rise of the Data Economy: Driving Value through Internet of Things Data Monetization". W: *IBM Corporation: Somers, NY, USA* (2016).

[137] Alexander Osterwalder i Yves Pigneur. *Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers*. Wiley, 2010.

[138] Sinno Jialin Pan, Qiang Yang i inni. "A survey on transfer learning". W: *IEEE Transactions on knowledge and data engineering* 22.10 (2010), strony 1345–1359.

[139] Amir Parssian, Sumit Sarkar i Varghese S Jacob. "Assessing data quality for information products: impact of selection, projection, and Cartesian product". W: *Management Science* 50.7 (2004), strony 967–982.

[140] Gerard Pasterkamp i inni. "Citation frequency: A biased measure of research impact significantly influenced by the geographical origin of research articles". W: *Scientometrics* 70.1 (2007), strony 153–165.

[141] Markus Perkmann i Henri Schildt. "Open data partnerships between firms and universities:Theroleofboundaryorganizations".W:*ResearchPolicy*44.5(2015), strony 1133–1143.

[142] Frank Piller i Joel West. "Firms, Users, and Innovation. An Interactive Model of Coupled Open Innovation". W: *New Frontiers in Open Innovation*. Redagował Henry Chesbrough, Wim Vanhaverbeke i Joel West. Oxford University Press, 2014. : 978-0-19-968246-1.

[143] Alessandro Piscopo. *Wikidata quality-A data consumers perspective*. 2017. : https://commons.wikimedia.org/wiki/File : Wikidata_quality-A_data_consumers_perspective.pdf (dostęp 12.02.2019).

[144] G.P. Podshivalenko. "Investicionnyj klimat i investicionnaja privlekatelnost". W: *Finansovaja analitika: problemy i reshenija* 15 (2010).

[145] Polska Wikipedia. *Pomoc:Przestrzeń nazw*. 2019. : https://pl.wikipedia.org/wiki/Pomoc:Przestrze%C5%84_nazw (dostęp 03.02.2019).

[146] Koduj Dla Polski. *Zbiory danych rowerowych dostępne w Polsce*. 2018. : https://kodujdlapolski.pl/zbiory-danych-rowerowychsa-dostepne-polsce/ (dostęp 08.01.2019).

[147] Joshua CC Pun i Frederick H Lochovsky. "Ranking Search Results by Web Quality Dimensions." W: *J. Web Eng.* 3.3-4 (2004), strony 216–235.

[148] PWC QUT Chair in Digital Economy. *RETAIL 5.0: CHECK-OUT THE FUTURE*. 2019. : https://chairdigitaleconomy.com.au/wp-content/ uploads/2018/04/Retail-5.0-Check-out-the-Future.pdf (dostęp 17.09.2019).

[149] Tomás Sáez i Aidan Hogan. "Automatically Generating Wikipedia Info-boxes from Wikidata". W: *Companion Proceedings of the The Web Conference 2018*. WWW '18. Lyon, France: International World Wide Web Conferences Steering Committee, 2018, strony 1823–1830. : 978-1-4503-5640-4. : 10. 1145 / 3184558. 3191647. : https://doi.org/10.1145/3184558.3191647.

[150] Igbal Safarov, Albert Meijer i Stephan Grimmelikhuijsen. "Utilization of open government data: A systematic literature review of types, conditions, effects and users". W: *Information Polity* 22.1 (2017), strony 1–24.

[151] Takaya Saito i Marc Rehmsmeier. "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets". W: *PloS one* 10.3 (2015), e0118432.

[152] Benjamin Saunders i inni. "Saturation in qualitative research: exploring its conceptualization and operationalization." eng. W: *Quality & quantity* 52 (4 2018), strony 1893–1907.

[153] Markus Schaal i inni. "Information quality dimensions for the social web". W: *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*. ACM. 2012, strony 53–58.

[154] Max Schmachtenberg, Christian Bizer i Heiko Paulheim. "Adoption of the linked data best practices in different topical domains". W: *International Semantic Web Conference*. Springer. 2014, strony 245–260.

[155] C. Seale. "Grounding theory". W: *The Quality of Qualitative Research*. Redagował C. Seale. London: SAGE Publications Ltd, 1999, strony 87–105.

[156] Nagel Sebastian. *December 2018 crawl archive now available*. 2018. : http://commoncrawl.org/2018/12/december-2018-crawlarchive – now – available/ (dostęp 17.01.2019).

[157] Per O Seglen. "Citation rates and journal impact factors are not suitable for evaluation of research". W: *Acta Orthopaedica Scandinavica* 69.3 (1998), strony 224–229.

[158] Sejm RP. *Obwieszczenie Marszałka Sejmu Rzeczypospolitej Polskiej z dnia 29 czerwca 2018 r. w sprawie ogłoszenia jednolitego tekstu ustawy o dostępie do informacji publicznej*. 2018. : http://prawo.sejm.gov.pl/isap. nsf/DocDetails.xsp?id=WDU20180001330 (dostęp 01.02.2019).

[159] Wenyi Shang. "A Comparison of the Historical Entries in Wikipedia and Baidu Baike". W: *International Conference on Information*. Springer. 2018, strony 74–80.

[160] Aili Shen, Jianzhong Qi i Timothy Baldwin. "A Hybrid Model for Quality Assessment of Wikipedia Articles". W: *Proceedings of the Australasian Language Technology Association Workshop 2017*. 2017, strony 43–52.

[161] Andrej Vladimirovich Shvecov. "Nekotorye metodicheskie podhody k jekonometricheskomu modelirovaniju vlijanija bjudzhetnoj politiki na jekonomiku". W: *Teorija i praktika obshhestvennogo razvitija* 3 (2011).

[162] Richard Socher i inni. "Recursive deep models for semantic compositionality over a sentiment treebank". W: *Proceedings of the 2013 conference on empirical methods in natural language processing*. 2013, strony 1631–1642.

[163] Anne H Soukhanov. *Encarta world English dictionary*. St. Martin's Press, 1999.

[164] Jan Filip Staniłko i Robert Kroplewski. *Polityka rozwoju sztucznej inteligencji w Polsce na lata 2019-2027*. Warszawa, 2019.

[165] David Stuart i inni. *Practical Challenges for Researchers in Data Sharing. Springer Nature survey data*. Sprawozdanie techniczne. 2018. : http://dx.doi.org/https://doi.%20org/10.6084/m9.figshare.5975011.

[166] Besiki Stvilia, Abdullah Al-Faraj i Yong Jeong Yi. "Issues of cross-contextual information quality evaluation-The case of Arabic, English, and Korean Wikipedias". W: *Library and Information Science Research* 31.4 (2009), strony 232–239. : 0740-8188. : 10.1016/j.lisr.2009.07.005.

[167] Besiki Stvilia i inni. "Assessing information quality of a community-based encyclopedia". W: *Proc. ICIQ* (2005), strony 442–454.

[168] Besiki Stvilia i inni. "Assessing information quality of a community-based encyclopedia". W: *Proc. ICIQ* (2005), strony 442–454.

[169] Besiki Stvilia i inni. "A framework for information quality assessment". W: *Journal of the American society for information science and technology* 58.12 (2007), strony 1720–1733.

[170] Q. Su i P. Liu. "A Psycho-Lexical Approach to the Assessment of Information Quality on Wikipedia". W: *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. Tom 3. 2015, strony 184–187. : 10.1109/WI-IAT.2015.23.

[171] Synchronicity. *Projekt H2020 Synchronicity – prezentacja*. 2018. : https: //synchronicity-iot.eu/wp-content/uploads/2018/06/Cities-cards_v3_Web.pdf (dostęp 08.01.2019).

[172] Świat Nauki. *Sieć Semantyczna*. 2001. : http://home.agh.edu.pl/ ~gjn/wiki/_media/dydaktyka:siec_semantyczna.pdf (dostęp 27.01.2019).

[173] Mike Thelwall. "Using altmetrics to support research evaluation". W: *International Workshop on Altmetrics for Research Outputs Measurements and Scholarly Information Management*. Springer. 2018, strony 11–28.

[174] Lőrinc Thurnay i inni. "ADEQUATe: A Community-Driven Approach to Improve Open Data Quality". W: *International Conference on Business Information Systems*. Springer. 2018, strony 555–565.

[175] Maria LorenaTuballa i Michael LochinvarAbundo. "A review of the development of Smart Grid technologies". W: *Renewable and Sustainable Energy Reviews* 59 (2016), strony 710–725.

[176] William Tunstall-Pedoe. "True knowledge: Open-domain question answering using structured knowledge and inference". W: *AI Magazine* 31.3 (2010), strony 80–92.

[177] Mieke L Van Driel, Manfred Maier i Jan De Maeseneer. *Measuring the impact of family medicine research: scientific citations or societal impact?* 2007.

[178] Joaquin Vanschoren i inni. "OpenML: networked science in machine learning". W: *ACM SIGKDD Explorations Newsletter* 15.2 (2014), strony 49–60.

[179] Joaquin Vanschoren i inni. "OpenML: networked science in machine learning". W: *CoRR* abs/1407.7722 (2014). arXiv: 1407.7722. : http://arxiv. org/abs/1407.7722.

[180] Antonio Vetrò i inni. "Open data quality measurement framework: Definition and application to Open Government Data". W: *Government Information Quarterly* 33.2 (2016), strony 325–337.

[181] Antonio Vetrò i inni. "Open data quality measurement framework: Definition and application to Open Government Data". W: *Government Information Quarterly* 33.2 (2016), strony 325–337.

[182] Denny Vrandečić i Markus Krötzsch. "Wikidata: a free collaborative knowledgebase". W: *Communications of the ACM* 57.10 (2014), strony 78–85.

[183] Jane Wakefield. *Burger King advert sabotaged on Wikipedia*. 2017. : https://www.bbc.com/news/technology-39589013 (dostęp 08.10.2018).

[184] Nick Wallace i Daniel Castro. *The State of Data Innovation in the EU*. Sprawozdanie techniczne. Center for Data Innovation, 2017, strony 1–115. : http://www2.datainnovation.org/2017-data-innovationeu.pdf.

[185] Jonathan Stuart Ward i Adam Barker. "Undefined by data: a survey of big data definitions". W: *arXiv preprint arXiv:1309.5821* (2013).

[186] Morten Warncke-wang, Dan Cosley i John Riedl. "Tell Me More : An Actionable Quality Model for Wikipedia". W: *WikiSym 2013*. 2013, strony 1–10. : 50318525. : 10. 1145 / 2491055. 2491063.

[187] Web Data Commons. *RDFa, Microdata, Embedded JSON-LD, and Microformats Data Sets – November 2018*. 2019. : http://webdatacommons.org/ structureddata/2018-12/stats/stats.html (dostęp 27.01.2019).

[188] Web Foundation. *Open Data Barometer. Global report, 4th edition, May 2017.* Sprawozdanie techniczne. 2017. : https://opendatabarometer.org/4thedition/report/ (dostęp 16.01.2019).

[189] Krzysztof Węcel i Włodzimierz Lewoniewski. "Modelling the Quality of Attributes in Wikipedia Infoboxes". W: *Business Information Systems Workshops*. Redagował Witold Abramowicz. Tom 228. Lecture Notes in Business Information Processing. Springer International Publishing, 2015, strony 308–320. : 978-3-319-26761-6. : 10. 1007 / 978 – 3 – 319 – 26762 – 3 _ 27. : http://dx.doi.org/10.1007/978-3-319-26762-3_27.

[190] Łukasz Wiechetek i Marek Mędrek. "Wykorzystanie otwartych zbiorów danych i systemów inteligencji biznesowej w jednostkach samorządu terytorialnego na przykładzie systemu JST Finanse". W: *Annales Universitatis Mariae Curie-Skłodowska, sectio H–Oeconomia* 50.2 (2016), strona 145.

[191] Szymon Wieczorek, Dominik Filipiak i Agata Filipowska. "Semantic Image-Based Profiling of Users' Interests with Neural Networks". W: *Studies on the Semantic Web* 36.Emerging Topics in Semantic Technologies (2018), strony 179–190. : 1868-1158. : 10. 3233 / 978 – 1 – 61499 – 894 – 5 – 179. : http://doi.org/10.3233/978-1-61499-894-5-179.

[192] Wikidata. *Item quality*. 2019. : https://www.wikidata.org/wiki/ Wikidata:Item_quality (dostęp 11.02.2019).

[193] Wikimedia Foundation. *Financial Statements*. 2016. : https://upload.wikimedia.org/wikipedia/foundation/4/43/Wikimedia_Foundation_Audit_Report_-_FY15-16.pdf (dostęp 08.10.2018).

[194] Wikimedia Strategic Planning. *Wikipedia Quality – Definition of quality*. 2018. : https://strategy.wikimedia.org/wiki/Task_Force/ Wikipedia_Quality/Definition_of_quality (dostęp 08.10.2018).

[195] Wikipedia Meta-Wiki. *List of Wikimedia projects by size*. 2018. : https: //meta.wikimedia.org/wiki/List_of_Wikimedia_projects_ by_size (dostęp 08.10.2018).

[196] Wikipedia Meta-Wiki. *List of Wikipedias*. 2018. : https://meta.wikimedia.org/wiki/List_of_Wikipedias (dostęp 08.10.2018).

[197] WikiStats. *List of largest Mediawikis*. 2018. : http://wikistats.wmflabs.org/largest_html.php (dostęp 08.10.2018).

[198] Dennis M Wilkinson i Bernardo a Huberman. "Cooperation and quality in wikipedia". W: *Proceedings of the 2007 international symposium on Wikis WikiSym 07* (2007), strony 157–164. : 10.1145/1296951.1296968. : http://portal.acm.org/citation.cfm?doid=1296951.1296968.

[199] Kewen Wu i inni. "Mining the factors affecting the quality of Wikipedia articles". W: *Information Science and Management Engineering (ISME), 2010 International Conference of*. Tom 1. IEEE. 2010, strony 343–346.

[200] Yanxiang Xu i Tiejian Luo. "Measuring article quality in Wikipedia: Lexical clue model". W: *IEEE Symposium on Web Society* 19 (2011), strony 141–146. : 21586985. : 10.1109/SWS.2011.6101286.

[201] Eti Yaari, Shifra Baruchson-Arbib i Judit Bar-Ilan. "Information quality assessment of community generated content: A user study of Wikipedia". W: *Journal of Information Science* 37.5 (2011), strony 487–498.

[202] Amrapali Zaveri i inni. "Quality assessment for linked data: A survey". W: *Semantic Web* 7.1 (2016), strony 63–93.

[203] Shiyue Zhang i inni. "History-Based Article Quality Assessment on Wikipedia". W: *Big Data and Smart Computing (BigComp), 2018 IEEE International Conference on*. IEEE. 2018, strony 1–8.

[204] Anneke Zuiderwijk i inni. "Open data for competitive advantage: insights from open data use by companies". W: *Proceedings of the 16th Annual International Conference on Digital Government Research*. ACM. 2015, strony 79–88.

Spis tabel

Spis rysunków

RAPORT

NOWY SUROWIEC

OTWARTE **ZASOBY DANYCH** DLA POLSKIEJ GOSPODARKI

Działalność biznesowa w coraz mniejszym stopniu opiera się na tradycyjnych czynnikach produkcji, takich jak kapitał czy praca, a rośnie wartość czerpana z posiadania i umiejętnego przetwarzania danych. Dostęp do danych jest obecnie jednym z bardziej istotnych bodźców wpływających na rozwój przedsiębiorstw. Ważna jest również umiejętność ich wykorzystania, a do tego niezbędne jest budowanie kompetencji w zakresie *data science*.

W niniejszym raporcie dokonano przeglądu dostępnych zasobów danych, wskazując na różnorodność typów i źródeł pochodzenia danych. Oceniono ich przydatność do budowania usług i aplikacji.

Dla przedsiębiorców poszukujących inspiracji zaprezentowano również analizę przypadków ilustrujących różne scenariusze wykorzystania i przetwarzania danych. **Przedstawione w raporcie treści pozwalają odpowiedzieć na najistotniejsze pytania dotyczące kierunków rozwoju polskich przedsiębiorstw najintensywniej pracujących z danymi**.

MINISTERSTWO ROZWOJU