Wtórne wykorzystanie danych administracyjnych w statystyce publicznej – obecnie i w przyszłości.

Ekspertyza opracowana na zlecenie

Polskiej Agencja Rozwoju Przedsiębiorczości[1].

Autor:

Dominika Rogalińska

Warszawa, marzec 2022 r.

[1] Wnioski zawarte w dokumencie nie stanowią oficjalnego stanowiska Głównego Urzędu Statystycznego.

Spis treści

1. Dane administracyjne i ich rola w statystyce publicznej

1.1. Transformacja badań statystycznych w kierunku systemu opartego na rejestrach

Służby statystyki publicznej w coraz większym stopniu wykorzystują źródła administracyjne – rejestry urzędowe, systemy informacyjne administracji publicznej, co spowodowane jest m.in.: dążeniem do redukcji obowiązków sprawozdawczych nakładanych na respondentów; ekonomiczną presją dotyczącą obniżania kosztów produkcji statystycznej; dostępem do tych źródeł będących wynikiem umów i odpowiednich regulacji prawnych, ale także rozwiązań IT i postępem technologicznym, który daje możliwość przetwarzania przez statystyków dużych wolumenów danych (stat.gov.pl, 2022). To jednocześnie odpowiedź na nowe potrzeby informacyjne polityków, decydentów i mieszkańców w zakresie tworzenia i realizowania polityk publicznych (w tym ich monitoringu i ewaluacji).

W badaniach realizowanych przez GUS, źródła administracyjne wykorzystuje się m.in. w celu uzupełnienia badań prowadzonych tradycyjnymi metodami (dane ze sprawozdań statystycznych, spisów), ich zastąpienia, bądź wykorzystania w nowych badaniach statystycznych, które dotychczas nie były przedmiotem prac statystyki publicznej (stat.gov.pl, 2022). W okresie pandemii Covid-19 krajowe urzędy statystyczne korzystały ze źródeł administracyjnych również w celu zrekompensowania luki w danych gromadzonych tradycyjnymi sposobami oraz aby przyspieszyć ich udostępnianie odbiorcom zewnętrznym. Źródła administracyjne nie zawsze były używane przez statystykę publiczną z jednakową intensywnością, co wiązało się z ograniczeniami natury prawnej, technicznej i informatycznej. W latach 90. zastosowanie danych administracyjnych w badaniach statystycznych miało charakter incydentalny i było związane głównie z wykorzystaniem rejestru PESEL na potrzeby realizacji badań demograficznych. Możliwość szerszego użycia danych administracyjnych pojawiła się wraz z rozwojem informatyki i odejściem od dokumentacji papierowej na rzecz elektronicznej.

W latach 2010-2021 nastąpił intensywny rozwój wykorzystania danych administracyjnych. Było to możliwe dzięki wdrożeniu, na potrzeby przeprowadzenia Powszechnego Spisu Rolnego 2010 (PSR) i Narodowego Spisu Powszechnego Ludności i Mieszkań 2011 (NSP), nowego środowiska informatycznego służącego przetwarzaniu dużych zbiorów danych (Operacyjnej Bazy Mikrodanych) oraz środowiska służącego do analizy statystycznej tego typu danych (Analitycznej Bazy Mikrodanych). Zarówno w PSR, jak i w NSP, korzystano z danych pochodzących ze źródeł administracyjnych i poza administracyjnych, a informacje od respondentów pozyskiwano przez internet (samospis internetowy) oraz z wywiadów realizowanych w ramach badania reprezentacyjnego lub pełnego (PSR).

W spisie powszechnym w 2021 r. w znacznie większym stopniu wykorzystano dane z różnych rejestrów administracyjnych i poza administracyjnych (np. rejestry PESEL, ZUS, NFZ, REGON, rejestry budynków i mieszkań). Dane od poszczególnych osób pozyskane zostały poprzez: obowiązkowy samospis internetowy (aplikacja na stronie internetowej GUS), wywiad telefoniczny albo wywiad bezpośredni, wspomagane komputerowo lub urządzeniem mobilnym (NSP 2011, 2021).

Wzrost wykorzystania źródeł administracyjnych był możliwy z uwagi na równoległe wprowadzanie przez statystykę publiczną odpowiednich rozwiązań organizacyjnych i legislacyjnych. Nowelizacja ustawy o statystyce publicznej z dnia 9 kwietnia 2015 r.[2] zapewniła m.in.: szeroki dostęp statystyki publicznej do danych administracyjnych, zwiększenie uprawnień Prezesa GUS do zgłaszania gestorom wniosków dotyczących zawartości informacyjnej i określenia wymogów dotyczących jakości rejestrów, co pozwoliło na zapewnienie stosowania w nich standardów identyfikacyjnych, klasyfikacyjnych i definicyjnych[3]. Wprowadzono też szczegółowe regulacje dotyczące przetwarzania danych osobowych, co umożliwiło szersze wykorzystanie tych rejestrów administracyjnych, które zawierają dane tego typu.

Instytucjonalnie bardzo ważnym etapem w całym procesie było utworzenie w Urzędzie Statystycznym w Warszawie, a następnie w GUS, komórki organizacyjnej specjalizującej się w przekształcaniu danych administracyjnych w dane statystyczne oraz ich integrację ze zbiorami statystyki publicznej. Podjęcie działań w statystyce publicznej było możliwe dzięki rozbudowie całego systemu i infrastruktury do przetwarzania informacji, od użycia, do budowy i aktualizacji wykazów. Utworzone zostało w Centrum Informatyki Statystycznej GUS odpowiednie środowisko sprzętowo-narzędziowe do przetwarzania danych z rejestrów urzędowych.
W 2021 r. GUS pozyskał 438 zbiorów danych z rejestrów urzędowych i systemów informacyjnych administracji publicznej. Pozyskane zbiory zostały wykorzystane w 123 badaniach objętych programem badań statystycznych statystyki publicznej (PBSSP) na rok 2021. Dane do tych badań pochodziły z 262 rejestrów[4] urzędowych i systemów informacyjnych administracji publicznej prowadzonych przez 80 jednostek administracji publicznej (stat.gov.pl, 2022).

[2] Ustawa z dnia 9 kwietnia 2015 r. o zmianie ustawy o statystyce publicznej oraz niektórych innych ustaw (Dz. U. poz. 855).

[3] Art. 13. Organy administracji publicznej, Zakład Ubezpieczeń Społecznych, Narodowy Fundusz Zdrowia, Komisja Nadzoru Finansowego, a także inne państwowe lub samorządowe osoby prawne, organy rejestrowe oraz inne podmioty prowadzące rejestry urzędowe lub niepubliczne systemy informacyjne, przekazują lub udostępniają nieodpłatnie służbom statystyki publicznej zgromadzone dane w szczegółowym zakresie, postaci i terminach, określonych w programie badań statystycznych statystyki publicznej, w szczególności w postaci zbiorów danych z systemów teleinformatycznych, w tym wyników pomiarów, danych monitoringu środowiska, a w przypadku braku systemu teleinformatycznego – w innej utrwalonej postaci.

[4] Rejestr może zawierać więcej niż jeden zbiór danych.

1.2. Wykorzystanie danych administracyjnych (definicje, charakterystyka)

Dane administracyjne to dane, które pochodzą z działania systemów administracji publicznej np. dane gromadzone przez organy rządowe i samorządowe do celów rejestracji, transakcji i prowadzenia ewidencji (Elias, 2014). Rozszerzoną definicję proponuje M. Woollard (2014), który wskazuje, że dane administracyjne to nie tylko informacje z rejestracji i ewidencji, ale są to również dane dotyczące świadczenia usług publicznych. Dane administracyjne mogą pochodzić z różnych systemów administracyjnych (np. dotyczących zgonów i urodzeń, edukacji, opieki zdrowotnej, podatków, partii politycznych, mieszkalnictwa, pojazdów), których gestorami są różne podmioty państwowe.

Kluczowym elementem danych administracyjnych są rejestry publiczne. Definicja zaproponowana przez T. Staweckiego (2007) zakłada, że rejestrem publicznym są zbiory informacji o osobach, rzeczach lub prawach, spełniające następujące warunki:

- zostały utworzone zgodnie z przepisami obowiązującego prawa,
- są prowadzone przez organ rejestrowy o charakterze publicznym,
- przyjęcie, utrwalenie, a następnie ujawnienie określonych w nim informacji następuje, co do zasady, w wyniku podjęcia przez organ rejestrowy decyzji,
- prowadzenie rejestru i ujawnianie w nim określonych informacji rodzi skutki prawne zarówno dla jednostki, której wpis dotyczy, jak i dla organów władzy publicznej,
- jest jawny, czyli dostęp do niego oprócz organu rejestrowego mają przynajmniej jednostki, których rejestr dotyczy, oraz inne organy władzy publicznej, a co do zasady szeroka kategoria publiczności. Jawność może być pełna lub ograniczona.

Zgodnie z Ustawą o statystyce publicznej z dnia 29 czerwca 1995 r., dane administracyjne zawierają informacje i dane zgromadzone w rejestrach urzędowych i systemach informacyjnych administracji publicznej. Art. 13 Ustawy opisuje systemy informacyjne administracji publicznej jako „*systemy zbierania, gromadzenia i przetwarzania informacji przez organy administracji publicznej, Zakład Ubezpieczeń Społecznych, Narodowy Fundusz Zdrowia, Komisję Nadzoru Finansowego, organy rejestrowe, inne państwowe lub samorządowe osoby prawne oraz inne podmioty prowadzące rejestry urzędowe*”.

W polskim systemie prawnym definicję rejestru publicznego sformułowano w ustawie z dnia 17 lutego 2005 r. o informatyzacji podmiotów realizujących zadania publiczne. W brzmieniu ustawy rejestr publiczny to *rejestr, ewidencja, wykaz, lista, spis albo inna forma ewidencji, służąca do realizacji zadań publicznych, prowadzone przez podmiot publiczny na podstawie odrębnych przepisów ustawowych*. W rozumieniu zapisów ustawy z dnia 29 czerwca 1995 r. o statystyce publicznej rejestrami urzędowymi są rejestry publiczne w rozumieniu ustawy *o informatyzacji działalności podmiotów realizujących zadania publiczne* (Dz. U. z 2021 r. poz. 2070) oraz *inne rejestry i ewidencje prowadzone na podstawie ustaw lub wydanych na ich podstawie aktów wykonawczych, zawierające informacje o podmiotach gospodarki*

narodowej i ich działalności, informacje o osobach fizycznych, ich życiu i sytuacji oraz dotyczące zjawisk, zdarzeń i obiektów.

Rejestry publiczne mają ograniczony zakres przedmiotowy w stosunku do danych administracyjnych i należy zachować ostrożność przy używaniu tych pojęć i nie stosować ich zamiennie. Podstawowym zadaniem rejestru jest – rejestracja – a więc ewidencja podmiotów, które funkcjonują w obszarze zainteresowania i *sensu stricto* nie gromadzą informacji o efektach funkcjonowania (wykorzystuje się je do zewidencjonowania jednostek podlegających obserwacji), tymczasem dane administracyjne są już bardziej zainteresowane tym obszarem.

Dane administracyjne są źródłem dużych i złożonych informacji ilościowych. Zakres i rodzaj wytwarzanych danych administracyjnych jest ściśle regulowany przez krajowe ustawodawstwo i/lub zasady określone przez organy administracji państwowej (Wallgren i Wallgren 2014). Ustawodawstwo daje organom państwowym prawo do zgłaszania, rejestrowania i wykorzystywania danych osób lub firm/instytucji. Dane administracyjne charakteryzują się dużym rozmiarem, mogą obejmować całą populację lub jej podzbiór. Przykładem są rejestry świadczeń socjalnych, które będą zawierały informacje jedynie o tych osobach, które wystąpiły o świadczenia.

Dane administracyjne są informacjami o wysokiej jakości, z reguły zgodne ze stanem faktycznym, ponieważ informacje wprowadzane do systemów urzędowych na ogół mają konsekwencje dla respondentów (np. pozytywne jak w przypadku wypłaty świadczeń socjalnych lub negatywne jak w przypadku kar i grzywien za składanie fałszywych deklaracji podatkowych) (Connelly i in., 2016). Zakres czasowy danych prowadzonych rejestrów może być bardzo długi. Zwykle obejmuje te same jednostki zapewniając im porównywalność w czasie. Niemniej, przy ich analizie należy mieć na uwadze zmieniające się uwarunkowania prawne i ograniczony wpływ statystyki na tworzenie prawa w zakresie danego rejestru oraz uwarunkowania organizacyjne związane z prowadzeniem rejestru czy rodzajem pozyskiwanych i przechowywanych w jego ramach informacji.

Rejestry urzędowe mogą być nieuporządkowane, a ich wykorzystanie na potrzeby analiz statystycznych będzie musiało być poprzedzone ich przygotowaniem – oczyszczeniem i uporządkowaniem danych do formatu wymaganego do analizy (Connelly i in., 2016). Dane administracyjne mają charakter danych wielowymiarowych, co osiągane jest zazwyczaj poprzez łączenie danych z różnych zbiorów w celu uzyskania pełnej informacji o badanym zjawisku (Elias, 2014).

1.3. Dane administracyjne – wykorzystanie w statystyce publicznej

Realizacja zadań statystyki publicznej w oczywisty sposób wiąże się z przetwarzaniem bardzo dużej ilości danych, które są wykorzystywane do celów wspomagania podejmowania decyzji istotnych z punktu widzenia polityk publicznych. GUS jest zaangażowany w działania zwiększające efektywność statystyki i jednocześnie zmniejszające obciążenie respondentów, poprzez ograniczanie pozyskiwania danych bezpośrednio od nich. Dlatego przy wyborze

źródeł danych, GUS w pierwszej kolejności uwzględnia dane z rejestrów urzędowych i systemów informacyjnych administracji publicznej.

Pozyskiwanie, gromadzenie, przetwarzanie i udostępnianie danych ze źródeł administracyjnych przez GUS oparte jest na zasadach i standardach Europejskiego Systemu Statystycznego. Przykładem stosowanych standardów są Wspólne Ramy Jakości Europejskiego Systemu Statystycznego, których podstawą jest Europejski Kodeks Praktyk Statystycznych (EKPS, 2017). Kodeks zawiera zasady dotyczące wykorzystania danych administracyjnych, w tym m.in. wskazuje na podstawy prawne do gromadzenia danych z wielu źródeł, tworzenia i rozpowszechniania statystyk oraz wskazuje, że państwa członkowskie mogą korzystać z mocy prawa z bezpłatnego dostępu do danych administracyjnym do celów statystycznych (*Zasada 2, EKPS, 2017)*[5]. EKSP wskazuje również, że w celu unikania nadmiernego obciążenia respondentów w jak największym stopniu krajowe urzędy statystyczne powinny wykorzystywać dostępne dane administracyjne (Zasada 9, *EKPS, 2017*)[6].

Adekwatne zapisy znajdują się w prawie polskim. Ustawa o statystyce publicznej w Art. 5a. Par. 2 wskazuje, że „*w celu realizacji zadań określonych w ustawie, w tym badań statystycznych, służby statystyki publicznej zbierają dane:*

1) z rejestrów urzędowych;

2) z systemów informacyjnych administracji publicznej;

3) z niepublicznych systemów informacyjnych;

4) od respondentów".

Natomiast Art. 13, który szczegółowo wskazuje instytucje zobowiązane do przekazywania danych, w ust. 5 zawiera zapis, że „*przy wyborze źródeł danych na potrzeby statystyki publicznej w pierwszej kolejności uwzględnia się dane administracyjne*".
Wykorzystanie danych administracyjnych ma duże znaczenie dla statystyki publicznej. Przyczynia się do doskonalenia procesów prowadzenia badań, w tym tworzenia operatów, a także poszerzania zakresu opracowań i analiz. Dane administracyjne wykorzystywane są w opracowaniu danych wynikowych z badań i zastępują informacje pozyskiwane bezpośrednio od respondentów, przyczyniając się tym samym do zmniejszenia ich obciążenia obowiązkami sprawozdawczymi. Ponadto, łączenie danych z różnych źródeł pozwala na tworzenie bogatych pod względem informacyjnym zbiorów, opisujących badane populacje i umożliwiających analizę związków pomiędzy danymi. Warunkiem powodzenia i efektywnego wykorzystania źródeł zewnętrznych w badaniach statystycznych jest bieżąca

[5] Organy statystyczne są wyraźnie upoważnione przepisami prawa do gromadzenia i dostępu do informacji z wielu źródeł danych dla celów statystyk europejskich. Organy administracji publicznej, przedsiębiorstwa, gospodarstwa domowe i wszyscy obywatele mogą mieć prawny obowiązek udzielenia dostępu do danych lub podania takich danych na potrzeby statystyk europejskich na żądanie organów statystycznych.

[6] Obciążenie respondentów jest współmierne do potrzeb użytkowników i nie jest nadmierne dla respondentów. Organy statystyczne monitorują obciążenie respondentów i określają docelowe poziomy jego redukcji w czasie.

i trwała współpraca partnerska z gestorami danych. Powinna się ona opierać się na wymianie wiedzy na temat nowych, dostępnych dla statystyki publicznej źródeł danych administracyjnych oraz form ich przekazywania.

Więcej korzyści i ograniczeń w wykorzystaniu danych administracyjnych dla statystyki publicznej zostało przedstawiona w *Tabeli 1.*

Tabela 1. Korzyści i ograniczenia wykorzystania danych administracyjnych dla statystyki publicznej

Korzyści
▪ redukcja obciążeń administracyjnych, w tym: zmniejszenie obciążenia respondentów obowiązkami związanymi z dostarczaniem informacji dla statystyki oraz ograniczeniem zakresu i częstotliwości zbieranych danych bezpośrednio od respondentów; ▪ poprawa efektywności badań prowadzonych przez statystykę publiczną poprzez obniżenie ich kosztów; ▪ ograniczenie albo wyeliminowanie błędów z badań ankietowych prowadzonych przez statystykę publiczną; ▪ zwiększenie dokładności danych wynikowych i oszacowań; ▪ zwiększenie elastyczności i szybkości reakcji statystyki publicznej na potrzeby użytkowników, a także dostarczanie informacji adekwatnych do ich potrzeb (zakresu informacyjnego, dezagregacji przestrzennej i dziedzinowej); ▪ możliwość odejścia od tradycyjnych spisów powszechnych ludności i mieszkań na rzecz spisów opartych na rejestrach i spisach łączonych (rezygnacja z badań masowych na rzecz corocznych aktualizacji danych spisowych); ▪ możliwości rezygnacji z tradycyjnego zbierania danych na formularzach statystycznych w całości lub w części; ▪ zwiększenie aktualności danych wynikowych oraz skrócenie czasu przetwarzania danych; ▪ zwiększanie zgodności metodologicznej systemów administracyjnych między sobą oraz z systemem statystyki publicznej; ▪ zwiększanie integralności systemów administracyjnych między sobą oraz z systemem statystyki publicznej; ▪ uzyskanie szerszego zakresu oficjalnych informacji statystycznych stanowiących podstawę do podejmowania decyzji.

Ograniczenia
▪ dane z rejestrów administracyjnych są zwykle przeznaczone do osiągnięcia określonego celu administracyjnego i charakteryzują się mniejszą elastycznością względem możliwości wykorzystania w innych celach; ▪ wykorzystanie danych administracyjnych musi zostać poprzedzone ich przygotowaniem do analizy, w tym przeglądem, czyszczeniem i dostosowaniem do oczekiwanego formatu; ▪ możliwość występowania błędów na etapie pozyskiwania danych, błędy raportowania, problemy z dopasowaniem rekordów; ▪ dla każdego zbioru danych administracyjnych istotne jest uwzględnienie definicji i metodologii, a także kwestii prawnych; ▪ częste zmiany w przepisach regulujących gromadzenie określonych informacji (np. zmiany w zasadach rozliczeń podatkowych); ▪ wykorzystanie danych administracyjnych przez podmioty nie będące właścicielami danych, odbywa się na warunkach określonych przez właścicieli danych (np. departamenty rządowe); ▪ systemy informatyczne organizacji wykorzystującej dane administracyjne muszą być zorganizowane tak, aby mogły obsługiwać formaty rejestrów administracyjnych, przy jednoczesnym zapewnieniu im odpowiedniego poziomu bezpieczeństwa.

Źródło: Opracowanie własne na podst.: stat.gov.pl (2022); Groen (2012); Conelly i in. (2016); George and Lee (2002); Crawford I Schultz (2014).

Głównymi gestorami źródeł administracyjnych, od których GUS pozyskuje dane są: Ministerstwo Finansów, Zakład Ubezpieczeń Społecznych, Kasa Rolniczego Ubezpieczenia Społecznego, Ministerstwo Edukacji i Nauki, Agencja Restrukturyzacji i Modernizacji Rolnictwa, Główny Inspektorat Ochrony Środowiska, Kancelaria Prezesa Rady Ministrów, Agencja Rozwoju Przemysłu S.A., Główny Inspektorat Weterynarii, Ministerstwo Rodziny i Polityki Społecznej, Ministerstwo Sprawiedliwości, Polski Fundusz Rozwoju S.A. GUS współpracuje również w zakresie pozyskania danych administracyjnych z jednostkami samorządu terytorialnego.

Dane administracyjne są wykorzystywane we wszystkich dziedzinach statystyki publicznej. Wyłącznie na systemach administracyjnych oparte są badania w obszarach ludność, procesy demograficzne w okresach między spisowych czy w obszarach oświata i wychowanie. W statystyce rachunków narodowych, statystyce środowiska, ochrony zdrowia, statystyce rynku pracy czy statystyce działalności przedsiębiorstw rejestry są wykorzystywane również w bardzo dużym zakresie.

Korzystanie z rejestrów urzędowych i systemów informacyjnych administracji publicznej przez statystykę ma też wymierne korzyści dla gestorów czyli dysponentów tych rejestrów. Statystyka publiczna analizuje pod względem jakościowym dane z rejestrów i współpracuje

z właścicielami tych rejestrów, w celu podnoszenia ich jakości. W ramach współpracy uzgadniana jest również m.in. zgodność metodologiczna, definicje oraz inne aspekty rejestrów. Niedoceniana jest również rola danych administracyjnych w zakresie zmian organizacyjnych w samej statystyce (np. wymuszają one informatyzację procesów badawczych).

Przy wykorzystaniu danych administracyjnych bardzo ważna jest akceptacja społeczna czyli utrzymanie wysokiego poziomu zaufania społecznego, poprzez wprowadzenie odpowiednich regulacji w zakresie bezpieczeństwa danych. To oznacza odpowiednie środki techniczne i organizacyjne, ograniczenie dostępu wewnętrznego do danych jednostkowych, identyfikowalnych z rejestrów oraz systemów i nieudostępnianie danych jednostkowych użytkownikom zewnętrznym. Statystyka posiada bezpieczne środowisko przetwarzania danych z rejestrów administracyjnych w szczególności danych jednostkowych identyfikowalnych.

Dla pełnej transparentności dotyczącej całego procesu wykorzystania danych administracyjnych przez statystykę publiczną stworzone zostało środowisko do pracy na zbiorach administracyjnych – Repozytorium Standardów Informacyjnych (RSI). Jest to portal internetowy GUS zawierający opisy rejestrów urzędowych i systemów informacyjnych administracji publicznej. Gromadzona jest tam wiedza na temat zasobów informacyjnych, których gestorami są inne jednostki administracji publicznej. Znajdują się tam również informacje dot. jakości systemów urzędowych i ich użyteczności dla statystyki publicznej oraz terminy aktualizacji systemów, co pozwala określić czy dane z tych rejestrów mogą zostać wykorzystane jako stałe źródło informacji. W RSI można znaleźć ogólne, a także szczegółowe informacje o każdym wykorzystywanym przez statystykę (i nie tylko) rejestrze i systemie. Zasilanie RSI przez gestorów systemów następuje na podstawie zapisów ustawy o statystyce publicznej. Dołączenie opisu kolejnego systemu informacyjnego do zasobów RSI następuje na wniosek Prezesa GUS skierowany do organów prowadzących systemy informacyjne administracji publicznej lub na wniosek organu prowadzącego system informacyjny administracji publicznej skierowany do Prezesa GUS. Dostęp do RSI jest powszechny i bezpłatny (RSI.stat.gov.pl, 2022).

Wprowadzenie informacji o zasobach administracji publicznej do RSI jest swoistą inwentaryzacją, która stwarza warunki do identyfikacji źródeł danych do badań statystycznych, w tym oczywiście spisów, do oceny jakości, w tym użyteczności źródeł danych z uwzględnieniem wymogów statystyki publicznej oraz – co jest bardzo ważne – do prowadzenia analiz zgodności metodologicznej rejestrów urzędowych i systemów informacyjnych administracji publicznej z systemem statystyki publicznej. System RSI pozwoli na lepsze wykorzystanie danych administracyjnych, poprawienie jakości systemu informacyjnego państwa, zwiększenie interoperacyjności systemów. Zasób informacyjny RSI to wiedza na temat 667 rejestrów (z blisko 80 urzędów administracji publicznej).

1.4. Zastosowanie big data w statystyce (definicje)

Statystyka przez dziesięciolecia z powodzeniem wypełniała swoją misję, wykorzystując tradycyjne metody pozyskiwania danych, a mianowicie spisy powszechne oraz badania statystyczne. Obecnie zwiększa się wykorzystanie danych z rejestrów administracyjnych ale nie są one ani jedynymi, ani najobszerniejszymi pozastatystycznymi źródłami danych wykorzystywanymi w badaniach statystycznych.

Oczywiste staje się, że wykorzystanie danych generowanych przez transformację cyfrową jest jedynym sposobem pomiaru pewnych zjawisk. Istnieje rosnąca przepaść między terminowością i szczegółowością dostarczania danych, które można osiągnąć za pomocą tradycyjnych źródeł danych. Jedyny sposób w jaki organy statystyczne mogą odpowiedzieć na te wyzwanie, pojawił się wraz z transformacją cyfrową i jest związany właśnie z big data.

Dotychczas nie została przyjęta jednoznaczna, powszechnie stosowana definicja pojęcia *big data*, co więcej de Goes (2013) posunął się nawet do zasugerowania, że termin big data jest zbyt niejasny i rozległy, aby miał znaczenie. Big data to dane niepróbkowane, charakteryzujące się tworzeniem baz danych ze źródeł elektronicznych, których podstawowy cel jest inny niż informacja statystyczna. Przez big data można rozumieć również dane wykorzystywane w innym celu, niż w tym, w jakim zostały utworzone (np. oferty pracy online wykorzystywane są w statystyce dot. rynku pracy) i/lub innowacyjne wielkoskalowe źródła danych ilościowych, które coraz częściej stają się dostępne dla celów badawczych (Conelly i in. 2016).

Big data można również zdefiniować poprzez dynamikę nowo pojawiających się zbiorów danych (wiele obserwacji z wieloma zmiennymi), charakter procesu zbierania danych (ciągły i automatyczny), formę gromadzonych danych (ustrukturyzowane i nieustrukturyzowane), źródła takich danych (publiczne i prywatne), „ziarnistość" danych (więcej zmiennych opisujących bardziej szczegółowe cechy osób, miejsc, wydarzeń, interakcji itd.) oraz skrócenie czasu między pozyskaniem danych, a ich analizą (Mergel, Rethemeyer i Isett, 2016).

Jedną z najczęściej przytaczanych definicji big data jest ta zaproponowana przez D. Laneya (2001), który opisuje big data poprzez atrybuty – tzw. koncepcja „3V", tj.: objętość – *volume* (duże wolumeny danych); różnorodność – *variety* (duża różnorodność danych w tym danych nieustrukturyzowanych); prędkość – *velocity* (duża zmienność i dynamika danych np. danych sensorycznych, strumieniowych, pochodzących z Internetu). W kolejnych podejściach definicyjnych zaproponowano rozszerzenie koncepcji „3V" o dodatkowe atrybuty (Villanova, 2014): wartość – *value* (duże znaczenie danych dla podejmowania decyzji biznesowych); wiarygodność – *veracity* (możliwość zaufania danym i korzystania z nich w sposób pewny, przy podejmowaniu kluczowych decyzji biznesowych).

Zgodnie z niektórymi poglądami koncepcja big data jest silnie związana ze skokową zmianą rodzajów zasobów danych, które stają się dostępne dla badaczy (Schroeder i Cowls 2014).

Można uznać, że koncepcja *big data* jest więc pewną odpowiedzią na problem przeciążenia informacyjnego (*information overload*), który choć głównie istnieje w kontekście biznesu, w praktyce dotyczy również sektora publicznego. Przeciążenie informacyjne to sytuacja, w której organizacja posiada dostęp do dużej ilości danych, ale nie ma możliwości ich przetworzenia i przeprowadzenia procesu wnioskowania (Wieczorkowski, Dałek 2013). Może więc nastąpić nadmiar informacji prawdziwej, lecz w praktyce nieprzydatnej ze względu na brak możliwości jej wykorzystania.

1.5. Big data – jako źródło danych statystycznych

Ze względu na rosnące oczekiwania i potrzeby odbiorców informacji statystycznej oraz doskonalsze technologie informacyjne, coraz większą uwagę przywiązuje się do wykorzystania big data przez statystykę publiczną. Wyzwaniem dla szerokiego wykorzystania tych zbiorów w produkcji statystycznej są m.in. kwestie prawne, dostępność danych, kwestie sposobu ich przechowywania, zagadnienia związane z konieczności zachowania tajemnicy statystycznej, czy – co równie istotne – problemy dotyczące obciążenia, będącego konsekwencją dużych wolumenów danych.

Big data mają duży potencjał do wykorzystania w administracji publicznej i usługach publicznych. W raporcie TechAmerica Foundation (2012) wskazano następujące możliwości zastosowania big data w usługach publicznych:

- poprawa jakości i efektywności funkcjonowania służby zdrowia,
- wczesne wykrywanie zagrożeń epidemiologicznych i sanitarnych,
- prognozowanie pogody i przewidywanie klęsk żywiołowych,
- wspomaganie zarządzania transportem,
- monitorowanie i podnoszenie jakości systemu edukacji,
- zarządzanie bezpieczeństwem w cyberprzestrzeni,
- wykrywanie nieprawidłowości w rozliczeniach podatkowych,
- monitorowanie rynku pracy i przeciwdziałanie bezrobociu.

Korzyści i ograniczenia stojące przed wykorzystaniem big data w statystyce zostały omówione w *Tabeli 2*.

Tabela 2. Korzyści i ograniczenia stojące przed wykorzystaniem big data w statystyce publicznej

Korzyści
▪ uzyskanie nowych informacji, które dotychczas były niedostępne w statystyce publicznej; ▪ możliwość wykorzystania big data jako zmiennych pomocniczych w prowadzonych badaniach (np. statystyce małych obszarów); ▪ poprawa efektywności wybranych badań prowadzonych przez statystykę publiczną poprzez obniżenie ich kosztów;

▪ możliwość zastąpienia, uzupełnienia lub poprawy istniejących zbiorów danych; ▪ redukcja obciążenia respondentów poprzez wykorzystanie już dostępnych danych; ▪ możliwość udostępniania wybranych danych w czasie rzeczywistym lub zbliżonym do czasu rzeczywistego; ▪ automatyzacja procesu pozyskiwania, przetwarzania i analizowania danych big data; ▪ nowe metody wizualizacji dużych zbiorów danych i wnioskowania statystycznego; ▪ zwiększenie aktualności danych wynikowych oraz skrócenie czasu przetwarzania danych.
Ograniczenia
▪ możliwy problem z dostępem do danych – „producentami" i „dostawcami" dużych zbiorów danych są jednostki sektora prywatnego; ▪ regulacje prawne nie są dostosowane do zmieniających się uwarunkowań organizacyjnych i technologicznych pozyskiwania, przetwarzania i udostępniania big data (np. kwestie ochrony prywatności i ochrony własności intelektualnej); ▪ big data na ogół nie spełniają wymagań metodologicznych statystyki publicznej (np. w kontekście definicji stosowanych przez statystykę publiczną); ▪ zróżnicowana jakość danych (np. błędy nielosowe na poziomie jednostki oraz źródła, pomiar w różnych odstępach czasu, dane nieustrukturyzowane); ▪ ograniczone możliwości integracji big data z istniejącymi źródłami statystycznymi (np. brak wspólnych identyfikatorów, różnej wielkości populacje); ▪ ograniczone możliwości wykorzystania istniejących struktur bazodanowych oraz metod stosowanych w statystyce publicznej; ▪ zróżnicowana jakość danych z punktu widzenia ich dokładności, przydatności, porównywalności, spójności, terminowości i punktualności; ▪ ograniczona możliwość zapewnienia pokrycia informacyjnego dla subpopulacji, dla których standardowo publikowane są dane w statystyce publicznej oraz związane z tym problemy identyfikacji podstawowych charakterystyk demograficznych.

Źródło: Opracowanie własne na podst.: Beręsewicz, Szymkowiak (2015); Conelly i in. (2016); George and Lee (2002).

1.6. Dane administracyjne a big data

Zgodnie z przytoczonymi definicjami big data, istnieje kilka głównych atrybutów, którymi możemy charakteryzować dane tego typu, w tym m. in. objętość, różnorodność, dynamika, wartość, wiarygodność (Villanova, 2014). Niektóre rodzaje danych mogą zawierać wiele z cech, które będą uważane za spełniające definicję dużych zbiorów danych (np. objętość/rozmiar, różnorodność i dynamika). Z kolei inne rodzaje dużych zbiorów danych mogą posiadać inny zestaw cech lub tylko jedną z nich, ale nadal mogą być uważane za dane typu big data (Kitchin 2014 a, b).

Nie wszystkie zasoby big data będą równie duże, nie wszystkie będą zapewniać szybką dostępność danych w czasie rzeczywistym i nie wszystkie będą zawierać równie szeroki zakres informacji. Wspólną cechą zasobów big data jest to, że są one danymi znalezionymi, a nie utworzonymi (Beręsewicz, Szymkowiak 2015). W *Tabeli 3* dokonano porównania między danymi administracyjnymi a big data w oparciu o wybrane atrybuty big data.

Tabela 3. Dane administracyjne a big data – analiza porównawcza

Atrybut	Dane administracyjne	Big data
Wolumen danych	Rejestry administracyjne są z definicji obszerne i obejmują wszystkie kwalifikujące się jednostki (ale nie zawsze[7]). Liczba jednostek rośnie wraz upływem czasu. Zbiory mogą być duże i liczyć od kilku tysięcy do kilkunastu milionów rekordów.	Zbiory danych mogą być bardzo duże i bardzo złożone. Przeważnie nie obejmują całej populacji, mogą obejmować różne jednostki. Przyrost danych może mieć charakter ciągły. Zbiory mogą liczyć kilkadziesiąt, a nawet kilkaset milionów rekordów.
Różnorodność	Rejestry administracji publicznej mogą obejmować dużą liczbę pozycji danych ustrukturyzo-wanych, zapewniając bogate źródło analiz w wielu wymiarach.	Big data mogą obejmować bardzo dużą liczbę pozycji danych. Są to dane zazwyczaj o charakterze nieustrukturyzowanym.[8]
Prędkość	Zbierane w sposób semi-systematyczny. Nie wszystkie z rejestrów są prowadzone w czasie rzeczywistym. Wiele z nich posiada osobne regulacje prawne związane z terminami ich aktualizacji lub uzupełnienia danych.	W zależności od źródeł danych istnieje duża zmienność i dynamika w pozyskiwaniu i przetwarzanych danych. Wiele danych może być pozyskiwanych i przetwarzanych w czasie zbliżonym do rzeczywistego.
Wartość (dla prowadzenia polityk publicznych)	Dane zostały utworzone przez organy administracji publicznej na potrzeby prowadzenia rejestrów, ewidencji.	Wartość jest uzależniona od „właściciela" i/lub „producenta" danych, w tym celów jakie dane mają realizować.
Wiarygodność	Duża wiarygodność danych. Zakres i rodzaj wynika z prawnie uregulowanej sprawozdawczości realizowanej za pośrednictwem organów administracji.	W zależności od wykorzystanego źródła danych, wiarygodność może być trudno weryfikowalna.

Źródło: Opracowanie własne.

[7] Dane administracyjne również mogą nie zawierać pełnej populacji – zależy jak zdefiniowana jest jednostka obserwacji np. rejestr płatników VAT nie obejmuje przedsiębiorstw, które nie rozliczają tego podatku i z punktu widzenia populacji przedsiębiorstw jest to zbiór niepełny.

[8] Dane big data mogą mieć również ustrukturyzowaną formę ze względu na stosowane oprogramowanie. Jeśli jakaś instytucja zewnętrzna posiada API to w praktyce posiada ustrukturyzowane dane (np. często w formacie JSON).

W przeprowadzonej analizie porównawczej można zauważyć, że w wielu przypadkach dane administracyjne i big data mogą mieć wspólne atrybuty. Wielkość przetwarzanych zbiorów może być do siebie zbliżona, jednak w przeciwieństwie do big data, źródła administracyjne charakteryzują się na ogół niższym wolumenem danych, liczonym w tysiącach lub kilku-kilkunastu milionach rekordów i odzwierciedlają jednostki statystyczne. Rejestry administracyjne są zazwyczaj dobrze zdefiniowane, ustrukturyzowane i przetwarzane są w standardowych bazach danych (np. SQL). Ich konstrukcja i sposób wykorzystania są charakterystyczne dla systemów ewidencyjnych. Z kolei big data mogą nie być do końca zdefiniowane, są zazwyczaj nieustrukturyzowane, a ich pozyskiwanie i przetwarzanie wymaga bardziej zaawansowanych rozwiązań programistycznych. Znaczna różnica zauważalna jest również w czasie pozyskiwania i przetwarzania danych. W przypadku danych administracyjnych ten czas jest regulowany prawnie, a w przypadku big data nie ma takiego obowiązku prawnego. Big data mogą być pozyskiwane i przetwarzane w czasie zbliżonym do rzeczywistego. Dane big data określane są również w literaturze mianem danych organicznych, które powstają wskutek ludzkiego działania, a nie zdefiniowanego i wystandaryzowanego badania statystycznego, dlatego ich wiarygodność również należy uznać za dyskusyjną, uzależnioną od źródła danych.

Dane administracyjne i big data mają cechy wspólne. Mogą być np. wykorzystywane dla pozyskania wiedzy na potrzeby realizacji polityk publicznych. Praktyczne wykorzystanie technologii związanych z analityką big data w statystyce publicznej należy uznać za działanie niezbędne dla sprostania współczesnym wyzwaniom informacyjnym szybko zmieniającej się rzeczywistości społecznej i gospodarczej, oraz dotrzymania kroku podmiotom sektora prywatnego prowadzącym działalność związaną z komercyjnym pozyskiwaniem, przetwarzaniem i udostępnianiem danych.

Wyzwaniem dla statystyki publicznej w tym zakresie jest wykorzystanie całego potencjału metodycznego statystyki, aby pokazać, że w zbiorach big data ukryta jest cenna wiedza, którą można i trzeba wykorzystać na potrzeby programowania i monitorowania krajowych, regionalnych i lokalnych polityk publicznych.

Dane prywatnych gestorów są dużą szansą i GUS aktywnie uczestniczy w eksploracji możliwości ich wykorzystania na potrzeby oficjalnej produkcji statystycznej. Statystyka publiczna rządzi się szeregiem wymogów w zakresie jakości danych, reprezentatywności, stabilności źródła, niezależności itd., a dane typu big data od prywatnych gestorów (np. dane internetowe) często nie spełniają ww. kryteriów. Stąd na poziomie europejskim i światowym podejmowane są intensywne prace nad ułatwieniem dostępu do tych danych m.in. Data Act (Digital Strategy, 2022), metodologią i kwestiami jakości. Na tym etapie, na płaszczyźnie statystyki publicznej, dane te wykorzystywane są głównie w charakterze statystyk eksperymentalnych, poza kilkoma przykładami zastosowania regularnego.

Dyskusja na temat wykorzystywania danych prywatnych gestorów w statystyce publicznej prowadzona jest również na poziomie europejskim, w ramach grupy zadaniowej ds. dostępu do danych prywatnych przez urzędy statystyczne. Prywatne firmy często dysponują danymi,

które mogłyby być przydatne w statystyce publicznej (np. dane telekomunikacyjne, które mogłyby zostać wykorzystane do wyszukiwania miejsc o słabej infrastrukturze, dużym obciążeniu sieci, itp.). Istnieje potrzeba rozszerzenia roli urzędów statystycznych w kierunku takim, aby stały się opiekunami ładu informacyjnego (ang. *data steward*), czyli nie tylko zbierały dane i produkowały na ich podstawie pewne wskaźniki, ale też dbały o jakość i bezpieczeństwo całego ekosystemu danych (taką rolę pełni już Urząd Statystyczny w Irlandii). GUS powinien być kluczowym aktorem w trwającym procesie tworzenia systemu informacyjnego infrastruktury państwa, wykorzystującego potencjał danych big data.

GUS podejmuje aktywne działania związane z badaniem możliwości wdrożenia i wprowadzaniem do produkcji statystycznej big data. Najbardziej zaawansowane prace wdrożeniowe lub przedwdrożeniowe są prowadzone w zakresie: wykorzystania zobrazowań satelitarnych na potrzeby statystyki rolnictwa; Systemu Automatycznej Identyfikacji (AIS, ang. *Automatic Identification Systemu*) oraz Elektronicznego Systemu Poboru Opłat (e-TOLL) w statystyce transportu morskiego i drogowego. Równolegle realizowane są badania dotyczące możliwości wdrożenia do produkcji statystycznej danych skanowanych[9] z sieci handlowych i danych scrapowanych[10] ze stron internetowych na potrzeby statystyki cen, oraz w ograniczonym zakresie danych internetowych na potrzeby uzupełnienia badania dot. rynku pracy, informacji o przedsiębiorstwach, rynku nieruchomości i turystyki.

Zobrazowania satelitarne wykorzystywane w badaniach statystyki rolnictwa pozwalają na identyfikację i monitorowanie upraw rolnych oraz ocenę wpływu zjawisk ekstremalnych takich jak: powódź, susza, przymrozki, podtopienia, itp. na stan upraw w okresie wegetacji. Wykorzystywane dane satelitarne pozyskiwane są bezpłatnie z europejskiego programu COPERNICUS. Zastosowane rozwiązania umożliwiają m.in. szybsze pozyskanie i prezentację danych związanych z uprawami, na niskim poziomie agregacji i w ujęciu przestrzennym, z pominięciem pracochłonnego przygotowania i przeprowadzania badania ankietowego.

Dane AIS i e-TOLL wykorzystywane w statystyce transportu pozwoliły wprowadzić nowe statystyki, szacowane w oparciu o opracowane modele, algorytmy i nowoczesne metody przetwarzania danych strumieniowych. Dane pozyskane z systemów mogą zostać wykorzystane do zarządzania systemami transportowymi oraz planowania i kształtowania polityki transportowej kraju. Wdrożenie nowych źródeł danych do produkcji statystycznej pozwoliło na uzyskanie szerszego zakresu informacji wynikowych, udostępnianych w krótkim czasie od ich pozyskania oraz wyliczenie wcześniej niedostępnych dla odbiorców wskaźników dotyczących: natężenia ruchu, pracy przewozowej, emisji zanieczyszczeń w transporcie.

[9] Dane skanowane (scanner data) to dane z sieci handlowych o dobrach konsumpcyjnych, uzyskane dzięki skanowaniu ich kodów kreskowych w punktach sprzedaży (Dane skanowane w pomiarze CPI, 2019).

[10] Webscraping to wydobywanie (ang. scrape) informacji z witryn internetowych (ang. web) z wykorzystaniem odpowiednich programów komputerowych. Struktura i zawartość strony internetowej są zakodowane w Hypertext Markup Language (HTML). Oprogramowanie scrapujące rozumie HTML, jest je w stanie analizować i wydobywać z niego informacje. Oprogramowanie może wyodrębniać określone pola informacji lub pobierać dokumenty, do których prowadzą linki na stronie (Nature.com, 2020).

Włączenie nowych źródeł do produkcji statystycznej pozwala na zmniejszenie obciążenia związanego z realizacją obowiązku sprawozdawczego, ograniczenie nakładu pracy i czasu niezbędnego do wypełniania formularzy, a także poprawę jakości i kompletności informacji na temat pojazdów i statków.

W fazie przygotowawczej do wdrożenia w produkcji statystycznej, w zakresie statystyki cen, są dane skanowane z sieci handlowych oraz scrapowane ze stron internetowych. Od stycznia 2021 r. rozpoczęto regularne pobieranie danych ze stron sklepów internetowych w zakresie określonego asortymentu (ceny ze sklepów są zbierane codziennie). Pozyskane dane są klasyfikowane do odpowiednich grup elementarnych Europejskiej Klasyfikacji Spożycia Indywidualnego według Celu (ECOICOP) wykorzystywanej w badaniu cen konsumpcyjnych. Istnieje możliwość dopasowania produktów obserwowanych w różnych momentach czasu (matching) co jest niezbędne w celu oceny zmiany ceny porównywanego produktu w czasie. Wykorzystane źródła danych oraz metody ich przetwarzania pozwalają na dokładniejszy pomiar zmian cen detalicznych towarów i usług, częściową automatyzację procesu produkcji statycznej (np. automatyczne przypisanie danych skanowanych przez ankieterów do odpowiednich kategorii produktów) oraz zapewnienie wysokiej jakości danych poprzez mechanizmy kontroli całego procesu. Wykorzystanie nowych źródeł danych pozwoli również na skrócenie czasu pozyskania i udostępnienia informacji.

Prowadzone badania eksperymentalne związane z wykorzystaniem danych scrapowanych ze stron internetowych dotyczą również pozyskiwania, przetwarzania, gromadzenia, i udostępniania danych pochodzących z Internetu, w tym obejmujących zagadnienia m.in.:

- internetowych ofert pracy – wykorzystanie danych pochodzących z portali, na których publikowane są ogłoszenia o pracę, na potrzeby prowadzenia statystyk związanych z rynkiem pracy, w tym zapotrzebowaniem na specjalistów w różnych dziedzinach;
- charakterystyki przedsiębiorstw – przetwarzanie informacji o przedsiębiorstwie, w celu ulepszania lub aktualizowania istniejących danych, takich jak obecność w Internecie, rodzaj działalności, informacje adresowe, struktura własności itp.;
- nieruchomości – pozyskiwanie informacji na temat charakterystyki sprzedawanych i wynajmowanych domów oraz mieszkań, w tym ich cen;
- statystyk dotyczących turystyki – wykorzystanie danych pochodzących z portali zawierających informacje o bazie noclegowej do aktualizacji kartotek obiektów noclegowych.

2. Zastosowanie i możliwości wykorzystania danych administracyjnych i big data dla celów badawczych i oceny polityk publicznych.

2.1. Wykorzystanie danych administracyjnych i big data w monitoringu i ewaluacji

Wraz z rozwojem technologii informacyjnych i nowych źródeł danych zwiększyły się również możliwości, jakie daje ewaluacja, rozumiana jako proces systematycznej oceny polityk publicznych. Informacje i dane są jednym z najcenniejszych zasobów wspierających rozwiązywanie problemów demograficznych, społecznych i gospodarczych. Istniejące zbiory w postaci danych administracyjnych i big data otwierają szerokie możliwości ich wykorzystywania do generowania nowej wiedzy, usprawniania podejmowania decyzji w procesie zarządczym oraz zwiększenia efektywności i przejrzystości w instytucjach publicznych. Tym bardziej, że coraz częściej pozyskanie danych staje się łatwiejsze. Ograniczeniami w wykorzystaniu nowych źródeł danych dla oceny realizacji polityk publicznych są kwestie związane z ich jakością, stabilnością oraz adekwatnością dla prowadzonych polityk.

W *Tabeli 4* dokonano oceny jakości i wiarygodności źródeł administracyjnych oraz big data, w oparciu o wybrane kryteria oceny źródeł, tj. *dostępności* (określenie czy dane są dostępne w momencie, kiedy są potrzebne właściwym odbiorcom), *dokładności* (określenie czy treść danych jest adekwatna do prowadzonej ewaluacji, czy dane precyzyjnie i dokładnie opisują określone zagadnienie podlegające ocenie), *kompletności* (określenie czy dane są wystarczające, aby móc je przetworzyć w konkretną wiedzę), *rzetelności* (określenie czy dane posiadają informacje potwierdzające prawdziwość przekazywanych danych), *terminowości i aktualności* (określenie czy dane są aktualne oraz czy częstotliwość aktualizacji treści jest zgodna z jej zawartością) (Nowakowski, 2015 za: Holmes, 1996; Mascott 2006).

Tabela 4. Ocena wartości źródeł danych na potrzeby monitoringu i ewaluacji

Wybrane kryterium oceny źródeł	Dane administracyjne	Big data
Dostępność	Danymi dysponują jedynie jednostki administracji publicznej. W wielu krajach służą one wyłącznie do użytku wewnętrznego instytucji odpowiedzialnych za ich pozyskiwanie i nie są	Danymi dysponują jednostki sektora prywatnego i publicznego. Dostęp do danych jest zróżnicowany i nie zawsze opiera się na przejrzystych zasadach ich udostępniania[11].

[11] Na przykładzie danych pozyskiwanych metodą *webscrapingu* - niezależnie od publicznego udostępniania danych na stronie internetowej, niektóre podmioty nie wyrażają zgody na ich pobieranie metodą *webscrapingu* lub stosują ograniczenia zakresu, wielkości, częstotliwości scrapowania.

	udostępniane zewnętrznie. Dostęp do danych regulowany jest prawnie bądź na podstawie doraźnych decyzji.	
Dokładność	Zakres i szczegółowość pozyskiwanych i przetwarzanych danych regulowana jest prawnie. Dane są ustrukturyzowane i gromadzone na poziomie jednostek.	Zakres i szczegółowość pozyskiwanych i przetwarzanych danych nie musi być regulowana prawnie. Dokładność danych jest zróżnicowania w zależności od źródła danych.
Kompletność	Dane obejmują całą obserwowaną populację, charakteryzują się wysokim poziomem kompletności. Instytucje publiczne dysponują bardzo szczegółowymi bazami danych administracyjnych, w różnych przekrojach czasowych. Dane różnych instytucji administracyjnych funkcjonują jako repozytoria obszernych ewidencji uczestników programów publicznych, podatników, beneficjentów pomocy społecznej, ubezpieczonych w publicznych systemach ubezpieczeń społecznych, stron podpisanych umów czy rejestrów osób bezrobotnych.	Dane raczej pochodzą z nieznanych populacji i mogą obejmować złożone i/lub nieznane próbki. Kompletność danych będzie zależna od źródła danych. Big data mogą być pozyskiwane m.in. ze źródeł internetowych, zobrazowań satelitarnych, operatorów sieci komórkowych, bądź pasywnie zbieranych danych z czujników.
Rzetelność	Dane z rejestrów charakteryzują się wysokim stopniem rzetelności ponieważ informacje wprowadzane do systemów na ogół mają konsekwencje prawne i finansowe dla jednostek. Niekiedy luki i niespójności w danych mogą wynikać z błędu ludzkiego przy ich wprowadzaniu do rejestru.	Dane pozyskiwane są z różnych źródeł zewnętrznych, w tym obejmują informacje wtórne. Możliwość weryfikacji poprawności jest ograniczona. Dane są pozyskiwane, przetwarzane i analizowane w sposób zautomatyzowany, co minimalizuje ryzyko błędu ludzkiego.
Terminowość i aktualność	Terminy przekazywania danych do rejestru oraz ich aktualizacja są regulowane prawnie.	Dane mogą być pozyskiwane i udostępnianie w bardzo krótkim czasie lub w czasie prawie rzeczywistym.

Źródło: Opracowanie własne.

Dane administracyjne mogą być szczególnie cenne dla opracowywania, monitorowania i oceny realizowanych polityk publicznych, ponieważ charakteryzują się wysoką jakością źródeł danych, a ich funkcjonowanie jest regulowane prawnie, co wpływa na ich wiarygodność, stabilność i ciągłość. Zapewniają zazwyczaj znacznie większą wielkość próby niż tradycyjnie realizowane badania statystyczne (Card i in., 2010) oraz mogą stanowić sposób na uzyskanie dostępu do informacji o tych grupach, które mogą być najmniej skłonne do udziału w badaniach ankieterskich. To potężne zasoby, zwłaszcza ze względu na wgląd w nierówności społeczne, zachowania ludzkie i skuteczność polityk społecznych, jakie mogą one oferować (Card i in., 2010; Einav i Levin, 2013).

Źródła big data również posiadają potencjalną wartość zastosowania w sektorze administracji publicznej oraz w ograniczonym zakresie mogą wspomagać monitorowanie i ewaluację realizowanych polityk państwa. W tym zakresie big data należy uznawać jako źródło danych, które może pełnić funkcję uzupełniającą względem danych administracyjnych. Techniki przetwarzania big data mogą być wykorzystane do kontroli wewnętrznej i nadzoru organów administracji publicznej. Zastosowanie nowych technologii i wykorzystanie sztucznej inteligencji do zarządzania dużymi bazami danych może zautomatyzować proces gromadzenia danych, poprawić ich jakość oraz zredukować liczbę błędów. To powinno przełożyć się na poprawę efektywności administracji publicznej oraz zmniejszenie obciążeń biurokratycznych (*Wykorzystywanie danych administracyjnych do ewaluacji polityk publicznych – wnioski i rekomendacje*, 2022).

2.2. Doświadczenia GUS w zakresie monitoringu i ewaluacji projektów rozwojowych

Jeśli przyjrzymy się obfitości danych gromadzonych i przetwarzanych przez statystykę publiczną, jej rola w ocenie polityk publicznych i interwencji publicznych wydaje się niemal naturalna. GUS ma bogate doświadczenia w wykorzystaniu danych zastanych na potrzeby ewaluacji interwencji publicznych – realizuje je od kilku lat we współpracy z ewaluatorami na zlecenie m.in. Ministerstwa Rozwoju, Ministerstwa Funduszy i Polityki Regionalnej, Polskiej Agencji Rozwoju Przedsiębiorczości, Narodowego Centrum Badań i Rozwoju. W badaniach tych wykorzystywane były m.in. dane z badań przedsiębiorstw i innowacyjności.

Statystyka publiczna przebyła długą drogę, aby zostać partnerem w procesie monitoringu i ewaluacji. Wszystko zaczęło się prawie 10 lat temu od próby wykorzystania firmowych mikrodanych do ewaluacji kontrfaktycznych. Na początku wyzwaniem było ustalenie efektywnych warunków współpracy, które z jednej strony umożliwiałyby korzystanie z wrażliwych mikrodanych, przy jednoczesnym zapewnieniu poufności danych i zgodności z innymi wymogami związanymi z użytkowaniem danych.

Od tego czasu GUS uczestniczył w szeregu ewaluacji interwencji i działań finansowanych przez UE w ramach polityki spójności, zawsze we współpracy z jednostkami zewnętrznymi,

zarówno krajowymi, jak i międzynarodowymi, takimi jak ministerstwa, agencje publiczne czy Bank Światowy.

Jednak z punktu widzenia statystyki publicznej i mając na uwadze bogactwo danych, którymi statystyka dysponuje, pojawia się pytanie o rzeczywistą skuteczność takiego podejścia do ewaluacji. Proces przeprowadzania ewaluacji efektów w określonym momencie, często nawet nie po zakończeniu wspieranych projektów, wydaje się dość nieskuteczny i niejednoznaczny, jeśli chodzi o dostarczenie wyczerpujących dowodów na potrzeby formułowania polityki[12]. Faktem jest, że w większości przypadków efekty interwencji rozciągają się na pewien okres czasu, więc efekty mierzone w danym momencie (zwykle sztucznie założone w dokumentach ewaluacyjnych) są niepełne i/lub niedoszacowane lub jeszcze nie pojawiły się. Zwykle ma to miejsce np. w przypadku wsparcia na innowacje, B+R, modernizację przedsiębiorstw itp., które mają długofalowe konsekwencje społeczno-gospodarcze. Jednocześnie obszary te cieszą się dużym zainteresowaniem rządów i polityk krajowych.

Jeśli dodamy do tego szybkie i dynamiczne zmiany w środowisku społeczno-gospodarczym, nie mówiąc już o niespodziewanych zdarzeniach kryzysowych, to wydaje się, że najskuteczniejszym podejściem do ewaluacji jest ewaluacja on-going (czyli ewaluacja interwencji przeprowadzanej w całym procesie, gdy zmiany są wdrażane) oraz ewaluacja ex-post w jakiś czas po zakończeniu programu.

Mówiąc o stworzeniu podstaw dla realnej polityki opartej na dowodach, należałoby zastosować zupełnie inne podejście do mierzenia efektów interwencji publicznych, a zamiast dołączać komponent ewaluacyjny do każdej konkretnej interwencji lub działania, powinno się pomyśleć o odrębnym „projekcie ewaluacyjnym", który umożliwiłby bieżący monitoring i ocenę wyników, które miały szansę urzeczywistnić się po zakończeniu wspieranych projektów.

Jeśli myślimy o zaangażowaniu statystyki publicznej w ewaluację i monitoring, niezwykle ważne jest, aby statystykę publiczną włączyć od samego początku procesu, czyli od etapu projektowania ewaluacji. Tylko w ten sposób możliwe będzie prawdziwe czerpanie z bogatego zasobu danych, ale także zrozumienie ograniczeń i wyzwań związanych z ich wykorzystaniem danych statystyki publicznej w ewaluacji. Tam gdzie statystyka była zaangażowana w działania ewaluacyjne na różne sposoby, w tym w sposób kompleksowy - od etapu projektowania ewaluacji we współpracy z agencją wdrażającą i ewaluatorami - tam rezultaty współpracy można uznać za satysfakcjonujące. Inne doświadczenia, mniej satysfakcjonujące, miały miejsce gdy zaangażowanie statystyki następowało dopiero pod koniec interwencji, z oczekiwaniem dostarczenia wskaźników wcześniej zaprojektowanych. Niestety tam, gdzie wybór danych odbył się bez udziału statystyki okazywało się, że

[12] Wynika to z dynamiki cyklu tworzenia polityk, gdzie nie przewiduje się przerwy pomiędzy wdrażanymi interwencjami (pozwalającej na uwidocznienie się pełni efektów). W praktyce, kontynuacja interwencji następuje w krótkim czasie po zakończeniu interwencji pierwotnej. Wpływa to na konieczność „wyprzedzenia" badania skuteczności.

wskaźniki przedstawione w planie ewaluacji, nie znajdywały zasięgu w danych statystyki publicznej.

Tak więc rozważania na temat wsparcia, jakie statystyka publiczna może zapewnić w ocenie i monitorowaniu, nie dotyczą wyłącznie nowych metod i źródeł danych, ponieważ metody są już dostępne (dostępnych jest na przykład kilka dobrze ugruntowanych metod kontrfaktycznych), a danych gromadzonych przez oficjalne statystyki w „tradycyjny" sposób jest już pod dostatkiem. W pierwszej kolejności powinniśmy skoncentrować się na tym, jak zapewnić stałą obecność statystyki publicznej na wszystkich etapach procesu ewaluacji, od początkowego etapu projektowania planów ewaluacji (np. poprzez przewidywanie w planie ewaluacji listy wskaźników, która byłaby konsultowana z GUS), poprzez budowę efektywnego, systemowego i kompleksowego podejścia, po generowanie rzetelnych oficjalnych danych statystycznych. Jeśli spojrzy się na statystykę publiczną, jako nieodłączną część procesu ewaluacji, jest ona w stanie otworzyć zupełnie nowy obszar integracji źródeł danych, pozwalający na pozyskiwanie kolejnych informacji i generowanie „szytych na miarę" statystyk do ewaluacji, a także dostarczać wiarygodnych informacji na potrzeby polityki opartej na dowodach.

2.3. Proces ewaluacji a big data

Do tej pory GUS nie korzystał ze źródeł big data w projektach ewaluacyjnych. Zgodnie z przytoczonymi w poprzednim rozdziale definicjami, używanie takich danych do oceny i monitorowania jest jeszcze przed nami, chociaż ten kierunek działań przygotowawczych został już rozpoczęty. Big data to zdecydowanie nowy trend i centralny punkt ekosystemu danych oraz obszar dużego zainteresowania statystyki publicznej. Wydaje się, że nie ma innej drogi dla oficjalnych statystyk, niż sięganie po big data, jeśli chcemy nadążać za dynamiką współczesnego świata i co może najważniejsze, odpowiadać na rosnące potrzeby użytkowników danych.

W GUS przeprowadzono wiele badań i prac eksperymentalnych, które są do pewnego stopnia możliwe do wdrożenia. Należy jednak podkreślić, że statystyka publiczna podlega pewnym reżimom jakościowym, metodologicznym, a także, o czym często się zapomina, trwałości. Oznacza to, że o ile wykorzystanie np. danych z telefonii komórkowej do produkcji statystyk otwiera nowe wymiary informacji, które można wygenerować, o tyle stabilny dostęp do tych danych prawie nie istnieje. GUS może uzyskać spektakularne jednorazowe wyniki na podstawie jednorazowego pozyskania danych od operatora sieci komórkowej, ale zasadniczym celem oficjalnych statystyk jest udostępnianie danych w długich szeregach czasowych i możliwość ich porównywania. Kluczowa jest więc stabilność źródła danych i wydaje się, że nie da się jej zapewnić bez nadrzędnych rozwiązań prawnych. Inną kwestią jest jakość danych i wyników statystycznych.

Istnieje szereg wyzwań związanych z wykorzystaniem danych pochodzących z innych źródeł niż statystyczne i rejestry administracyjne do oceny polityk publicznych. Na przykład, jeśli myślimy o ewaluacji wpływu z podejściem kontrfaktycznym, informacje wykorzystywane

podczas ewaluacji muszą być wyraźnie powiązane z podmiotami objętymi interwencją i jednostkami w grupie kontrolnej. Ponadto takie informacje są najczęściej objęte tajemnicą prawną i firmową oraz nie są dostępne w otwartych źródłach danych.

Biorąc to wszystko pod uwagę, GUS skupia się obecnie na wykorzystaniu danych administracyjnych do badań statystycznych i eksperymentuje z big data. Potencjał danych administracyjnych jest naprawdę imponujący i wykorzystanie ich na potrzeby ewaluacji stanowi kolejny etap w prowadzeniu badań ewaluacyjnych.

Patrząc w przyszłość, system danych administracyjnych w Polsce pozwala pozyskiwać coraz więcej informacji, łatwiejszych do przetworzenia i niedostępnych z innych źródeł. Z punktu widzenia statystyki publicznej, jeśli zastosujemy zaawansowane techniki przetwarzania i integracji danych, uzyskamy dane niezbędne do przeprowadzenia oceny pomocy publicznej. Podsumowując, wykorzystanie big data w GUS jest bardzo obiecujące i są obszary, w których może wzbogacić (ale nie zastąpić!) statystyki oficjalne.

2.4. Dane administracyjne w ewaluacji wsparcia przedsiębiorstw

W PBSSP prowadzonych jest corocznie ponad 100 różnego rodzaju badań przedsiębiorstw. Polska należy do grupy krajów, w których dostarczana jest bardzo bogata i zróżnicowana informacja z zakresu statystyki gospodarczej, co ma wpływ na znaczne obciążenia sprawozdawcze podmiotów gospodarczych. Dla redukcji tych obciążeń w statystyce publicznej prowadzone są działania związane z szerokim wykorzystaniem źródeł administracyjnych w celu poprawy jakości operatów do badań, imputacji danych, uogólniania wyników w badaniach reprezentacyjnych, zastępowania badań statystycznych danymi administracyjnymi i rozszerzenia możliwości analitycznych. Warto podkreślić, że niezależnie od wymagań stawianych przez statystykę publiczną, na przedsiębiorcach spoczywają też inne obowiązki administracyjne (np. przekazywanie informacji do ZUS czy urzędów skarbowych). Wypełniając zobowiązania statystyczne i administracyjne przedsiębiorcy są zmuszeni do dwukrotnego informowania o poziomie niektórych cech.

Stale prowadzona współpraca z Ministerstwem Finansów zaowocowała nowymi obszarami bardzo cennymi dla statystyki publicznej. Dotychczas otrzymywane dane z Ministerstwa Finansów z obszaru podatku dochodowego od osób prawnych (CIT) i osób fizycznych (PIT) oraz z Krajowej Ewidencji Podatników (CRP KEP) pozwalały na weryfikację i poprawę wyznaczania aktywności podmiotów, na potrzeby prowadzonych badań statystycznych, co było niezbędne do poprawy i optymalnego doboru podmiotów do badania. Było to ważne, na przykład, w reprezentacyjnym badaniu mikroprzedsiębiorstw, które pozwalało na polepszenie jakości operatu.
Nowe dane, które statystyka publiczna otrzymuje z Jednolitych Plików Kontrolnych VAT z części ewidencyjnej stanowią ważne źródło informacji, przede wszystkim w zakresie analizy przepływów między jednostkami prawnymi, tworzącymi jednostkę statystyczną

„przedsiębiorstwo”. Dane te są też wykorzystywane w badaniu statystyki strukturalnej oraz demografii przedsiębiorstw[13].

Niestety wykorzystanie danych administracyjnych na potrzeby ewaluacji programów wsparcia przedsiębiorstw to wyzwanie, które obecnie nie znajduje w pełni zastosowania (w przypadku badań współrealizowanych przez GUS). Przyczyną tego stanu rzeczy jest brak spójnych rozwiązań systemowych w zakresie udostępniania danych jednostkowych nieidentyfikowalnych.

Do tej pory na potrzeby ewaluacji programów wsparcia przedsiębiorstw wykorzystywane były głównie dane pochodzące z badań statystycznych statystyki publicznej. Wydaje się, że zastosowanie w tym celu danych pochodzących z rejestrów administracyjnych nie zmieni tego statusu i raczej będzie miało charakter marginalny. Za pewną przyczynę można przyjąć brak spójnych rozwiązań systemowych w zakresie udostępniania nieidentyfikowalnych danych jednostkowych (taka decyzja leży w gestii właściciela prowadzącego dany rejestr). Rozwiązaniem jest wykorzystanie bogatego zasobu danych rejestrowych gromadzonych przez GUS i we współpracy z GUS. Jednak to nowe podejście niesie za sobą wiele wyzwań. Dane administracyjne nie mogą zostać udostępnione w zakresie objętym tajemnicą statystyczną i ze względu na fakt, że GUS nie jest ich gestorem, są bardziej chronione niż dane statystyczne. Jedynie dane z rejestrów publicznych mogą zostać udostępnione.

Biorąc pod uwagę źródła informacji dla przedsiębiorców prowadzących działalność gospodarczą i porównując je z rejestrami statystycznymi możemy mieć do czynienia z różnymi jednostkami obserwacji, innym zakresem przedmiotowym zmiennych, sposobem aktualizacji, co w konsekwencji prowadzi do braku porównywalności danych z tych źródeł. Co ciekawe różnice definicyjne dotyczą również samego podmiotu badania, czyli przedsiębiorstwa. Mamy tu do czynienia z art. 10 ustawy o statystyce publicznej z dnia 29 czerwca 1995 r. versus art. 4 ustawy z dnia 6 marca 2018 r. Prawo przedsiębiorców. Statystyka publiczna posługuje się pojęciem *przedsiębiorstwo,* rejestry administracyjne – terminem *przedsiębiorca*, a ich definicja nie jest spójna.

Przyczyną niejednolitego podejścia do jednostek prawnych jest też cel w jakim dany rejestr został powołany. Rozbieżności np. dotyczące pojęcia „osoby fizycznej” czy „osoby prawnej” występują między systemem REGON a pozostałymi rejestrami administracyjnymi. Na pewno dużym wyzywaniem byłoby utworzenie bazy wynikowej dla przedsiębiorstw obejmującej wyszczególnione cechy z danych znajdujących się w wielu rozproszonych systemach i rejestrach.

Kierunek nakreślony przez Komisję Europejską w zakresie wykorzystania danych statystyki publicznej i danych administracyjnych dla potrzeb monitoringu i ewaluacji wsparcia

[13] Ze względu na uwarunkowania prawne (nie każde przedsiębiorstwo jest czynnym podatnikiem VAT, część podatników korzysta ze zwolnień podatkowych), nie dla całej populacji przedsiębiorstw objętych badaniami statystycznymi będą informacje w zbiorach VAT Ministerstwa Finansów.

przedsiębiorstw to słuszna teza. Jest to jednak kwestia złożona i wymagająca podjęcia szeregu zabiegów na poziomie legislacyjnym, metodologicznym i technicznym.

2.5. Dane administracyjne w ewaluacji rynku pracy

Dane ZUS oraz KRUS są jednymi z najważniejszych, które mogą jednocześnie dostarczyć informacji zarówno o miejscu zatrudnienia, jak i miejscu zamieszkania poszczególnych osób i to według szczegółowego podziału terytorialnego kraju. Przedsiębiorstwa i inne jednostki sprawozdawcze nie udostępniają tego typu informacji. Dane o zatrudnieniu przekazywane są do GUS zbiorczo i dotyczą całej jednostki sprawozdawczej, a nie każdego pracującego oddzielnie.

W ramach prac badawczych zrealizowanych przez GUS w programie badawczym „Statystyka dla polityki spójności", finansowanym ze środków Programu Operacyjnego Pomoc Techniczna[14], przeprowadzono identyfikację źródeł informacji możliwych do zastosowania w statystyce publicznej w obszarze rynku pracy wraz z ich wnikliwą analizą. W wyniku tych podejść eksperymentalnych dokonano oceny zgodności/rozbieżności w definicjach i zakresach podmiotowych oraz przedmiotowych potencjalnych źródeł danych, z dotychczas stosowanymi w statystyce.

Pozyskiwanie danych ze źródeł administracyjnych dla badań z obszaru rynku pracy pozwala szybciej dostarczać statystyk, które będą bardziej adekwatne do monitorowania zmieniającej się sytuacji na rynku pracy i dostosowywania działań gospodarczych oraz społecznych do tej zmiennej sytuacji. Co jest również istotne, dane administracyjne pozwalają przygotować statystyki odnoszące się do całej zbiorowości, a więc również podmiotów tych najmniejszych, dla których statystyka ma stosunkowo mało danych z badań statystycznych. Dane administracyjne pozwalają również agregować dane z zakresu rynku pracy o pracujących i wynagrodzeniach na niskim poziomie agregacji terytorialnej tj. do poziomu gminy chociażby według miejsca zamieszkania pracujących. To dodatkowa zaleta, ponieważ dane agregowane w sprawozdawczości odnoszą się do miejsca lokalizacji podmiotów, którymi nie są osoby zatrudnione. Dodatkowym atutem jest możliwość prezentowania danych według płci, wieku czy statusu zatrudnienia, jak również z wykorzystaniem miar pozycyjnych i z większą częstotliwością.

Wyzwania, które stoją przed statystyką, dotyczą podjęcia działań związanych z wykorzystaniem tych danych w regularnej produkcji statystycznej. Proces pracy na źródłach administracyjnych opiera się na doprowadzeniu do optymalnej zgodności pojęciowej w zakresie pracujących i wynagrodzeń.

[14] Badania „Opracowanie metodologii i oszacowanie danych dotyczących pracujących w gospodarce na poziomie powiatu (NTS 4)" oraz „Opracowanie metodologii i oszacowanie liczby pracujących w gospodarce narodowej według głównego miejsca pracy i miejsca zamieszkania na poziomie powiatów, stopy bezrobocia rejestrowanego na poziomie gmin oraz miar wynagrodzeń brutto na poziomie powiatów". Szersza informacja o badaniach eksperymentalnych znajduje się na stronie https://stat.gov.pl/statystyki-eksperymentalne/kapital-ludzki/.

Kolejne wyzwania dotyczą kontaktu z respondentem. Sprawozdawczość statystyki publicznej opiera się na kontakcie pomiędzy statystykiem a respondentem. W przypadku otrzymywania zbiorów danych od gestora, ten kontakt w zasadzie jest wyeliminowany i brakuje możliwości bezpośredniej weryfikacji. W związku z tym wyjaśnianie nietypowych sytuacji musi przebiegać inną drogą.

Następną ważną kwestią są zmiany prawne zachodzące w rejestrach, które za każdym razem, jeżeli zachodzą, muszą znaleźć odzwierciedlenie w zapisach przekazywanych do statystyki. I najważniejsze wyzwanie przed którym stoi GUS to sposób publikowania danych z administracyjnych źródeł danych z uwzględnieniem zachowania zasad tajemnicy statystycznej.

3. Wyzwania i możliwości związane z pracą na dużych zbiorach danych administracyjnych

3.1. Przygotowanie zbiorów danych z systemów administracyjnych do wykorzystania w statystyce

Rejestry administracyjne jako niewątpliwie bogate źródło informacji nie zostały przygotowane i zorganizowane do potrzeb statystyki, a do potrzeb gestorów. Całą sztuką jest więc odpowiednia transformacja tych danych, przygotowanych do innych celów niż statystyczne, tak, żeby były one przydatne statystyce.

Wykorzystanie danych administracyjnych można analizować m.in. z punktu widzenia prawa, jakości, techniki, standaryzacji, organizacji, kosztów i zasobów. Aspekty prawne dotyczą formalnego uregulowania dostępu do danych z systemów administracyjnych. Przykładem takiego uregulowania jest Ustawa z dnia 29 czerwca 1995 r. o statystyce publicznej (Dz.U. z 2022 r. poz. 459). Na podstawie ustawy powstaje zobowiązanie gestorów systemów administracyjnych do przekazywania danych. Informacje o zakresie, częstotliwości i terminie przekazywania danych zawarte są w PBSSP, uchwalanym co roku w formie rozporządzenia Rady Ministrów.

Aspekty jakościowe wynikają z należytego zaprojektowania zbioru i zapewnienia jego budowy w zgodzie z tym projektem. Jakość zbioru jest związana z eliminacją błędów zbioru wynikających ze sposobu wprowadzania danych i ich przetwarzania. Za zaprojektowanie zbioru odpowiada gestor, czyli osoba prawna prowadząca dane źródło, będąca jego właścicielem. Z aspektem jakościowym wiąże się problem poziomu agregacji danych. Dla statystyki publicznej priorytetem jest pozyskiwanie danych jednostkowych, identyfikowalnych. Jest to najniższy możliwy poziom agregacji umożliwiający dalsze, dowolne przekształcanie zbioru. Na etapie przygotowania danych do publikacji, jeśli zachodzi potrzeba ukrycia danych poszczególnych respondentów, następuje przekształcenie danych poprzez połączenie poszczególnych rekordów danych dla ich prezentacji w postaci agregatów. Proces

agregacji nie jest odwracalny – bez dodatkowych danych nie można dokonać podziału agregatów na dane jednostkowe.

Aspekty techniczne dotyczą formy dostępności danych, sposobu ich pozyskiwania i przetwarzania, w tym zapisu powiązania danych z informacją przestrzenną. Dla opracowania danych przestrzennych kluczowa jest dostępność danych w postaci elektronicznej. Informacja przestrzenna może być zapisana z różną dokładnością – od poziomu jednostki administracyjnej (kraj, gmina itd.), poprzez kod pocztowy, składowe adresu pocztowego do poziomu współrzędnych geograficznych jednoznacznie identyfikujących położenie punktu w przestrzeni. W tym przypadku preferowanym poziomem szczegółowości danych przestrzennych jest poziom punktu np. punktu adresowego. Możliwe jest nieodwracalne agregowanie przestrzenne danych np. poprzez przypisanie wszystkich punktów do określonych gmin.

Z aspektem technicznym wiąże się zastosowanie standardów identyfikacyjnych takich jak numer identyfikacyjny PESEL, NIP czy kod TERYT. Zastosowanie standardów radykalnie poprawia jakość danych przez ujednoznacznienie treści zapisów rejestru, ogranicza prawdopodobieństwo mylnego wprowadzenia danych dzięki zastosowaniu słowników i narzędzi sprawdzających poprawność wprowadzonych danych.

Organy administracji publicznej prowadzące systemy i rejestry administracyjne zobligowane są do stosowania w nich standardów klasyfikacyjnych i identyfikacyjnych – na przykład numeru identyfikacyjnego i oznaczeń kodowych przyjętych w krajowym rejestrze urzędowym podziału terytorialnego kraju TERYT. Mimo tego w systemach informacyjnych administracji publicznej widoczna jest duża dowolność zapisu nazw miejscowości oraz identyfikatorów adresowych ulic. Stosowane są nieurzędowe nazwy miejscowości lub własne nazwy, będące połączeniem nazw urzędowych, a nazwy ulic odbiegają od nazw nadanych w uchwałach rad gmin, co z kolei wymaga dodatkowych prac standaryzacyjnych.

Spójność klasyfikacyjna i identyfikacyjna jest kluczem do sprawnej wymiany danych pomiędzy rejestrami oraz ich integracji na potrzeby statystyczne. Identyczne zmienne pochodzące z różnych rejestrów mogą być jednak zakodowane na różne sposoby, co w praktyce utrudnia integrację danych z wielu źródeł i wyprowadzenie wskaźników dla pełnej zbiorowości. Tym samym wymaga szeregu prac zmierzających do zapewnienia spójności m.in. poprzez opracowanie jednolitego standardu kodowania zmiennych i opisania go w systemie metadanych. Poprawnie zbudowany system metadanych pozwala na szybkie uzyskanie informacji o zbiorach danych dotyczących konkretnej dziedziny, wyjaśnienie rozbieżności pomiędzy zmiennymi pochodzącymi z różnych źródeł, wyeliminowanie dublujących się informacji i ich harmonizację, a tym samym poprawę użyteczności i jakości danych. Przy wykorzystaniu danych pochodzących z różnych rejestrów administracyjnych, często można spotkać się z: redundancją danych, różnicami metodologicznymi, brakiem interoperacyjności pomiędzy rejestrami, czy brakiem spójności informacji zawartych w rejestrach. Stąd też, przed rozpoczęciem analiz pozwalających na ocenę jakości źródła i możliwości jego wykorzystania do celów statystycznych, dokonuje się szeregu procesów

umożliwiających w konsekwencji poprawę jakości danych. W ramach przekształceń m.in. usuwane są rekordy powtarzające się, następuje ujednolicenie nazw rekordów, a następnie walidacja, korekta i integracja. Proces przekształcenia danych jednostkowych umożliwia uzyskanie zbioru, w którym dane są spójne, a zmienne zawierają wartości wystandaryzowane i poprawne merytorycznie. Proces przeprowadzany jest na danym zbiorze jeden raz – na rzecz wszystkich dalszych opracowań. W przypadku modyfikacji zbioru pozyskiwanego od gestora, zmianie ulega również proces jego przetwarzania.

Aspekty organizacyjne dotyczą sposobu prowadzenia badań, w tym ich cykliczności. Podstawowe badania prowadzone są w cyklach rocznych. Wyróżniamy także badania krótkookresowe (np. w cyklach kwartalnych) oraz badania realizowane w cyklach kilkuletnich. Realizacja badania w określonym cyklu jest związana przede wszystkim ze zmiennością zjawiska. Zmienność zjawisk w czasie i przestrzeni wymaga odpowiedniej aktualizacji rejestrów. Z sytuacją idealną mamy do czynienia wtedy, gdy źródło jest aktualizowane w czasie rzeczywistym lub zbliżonym do rzeczywistego. Wówczas stan danych w rejestrze zmienia się z pomijalnie niewielkim opóźnieniem w stosunku do rzeczywistości. W takiej sytuacji możliwe jest pozyskanie danych ze źródła według stanu na określony moment np. na koniec roku.

Z odwrotną sytuacją mamy do czynienia w przypadku źródeł aktualizowanych fragmentarycznie np. poprzez weryfikację stanu danych w części jednostek administracyjnych w jednym roku. Fragmentaryczną aktualizację stosuje się głównie ze względu na dużą skalę i koszty prac. Przykładem źródła aktualizowanego w taki sposób jest Baza Danych Obiektów Topograficznych (BDOT). Taka forma ogranicza możliwości pozyskania danych do celów statystycznych, ponieważ dane dla całego kraju pozyskane w jednym momencie obrazują stan z różnych okresów i nie mogą być zestawione z danymi statystycznymi obrazującymi stan zjawiska obserwowany danego dnia np. 31 grudnia danego roku.

Rejestry mogą być prowadzone centralnie tj. przez jednego gestora lub w sposób rozproszony, kiedy gestorów jest wielu. Wykorzystanie zbiorów rozproszonych wymaga wcześniejszego ich scalenia. W bardziej złożonych sytuacjach proces zbierania danych może wymagać przygotowania różnych form i organizacji przekazywania danych przez gestorów.

Zatem efektywne wykorzystanie zasobów, zarówno tych administracyjnych, jak i tych niepublicznych, jest dla statystyki publicznej wyzwaniem zarówno w zakresie metodologicznym, jak i programistycznym. Wyzwanie w obszarze metodologii wiąże się przede wszystkim z wypracowaniem metod pozwalających na przekształcenie danych przechowywanych w rejestrach na ogół do celów ewidencyjnych, w dane statystyczne. Wyzwaniem jest również oprogramowanie wypracowanych metod w narzędziach do automatycznego przetwarzania danych. Do tego rodzaju działań potrzebna jest wielokierunkowa wiedza z zakresu statystyki, metodologii badań oraz pewnej specyfiki rejestrów. Rejestry powinny być odpowiednio przygotowywane, standaryzowane i wykorzystywane w pracach statystyki publicznej tak, aby proces korzystania z nich był

systemowy i odbywał się w ramach dedykowanego, powtarzalnego algorytmu. Założeniem jest więc efektywne wykorzystanie zasobów informacyjnych zawartych w rejestrach poprzez zintegrowanie wybranych zbiorów danych oraz stworzenie złożonego systemu umożliwiającego obserwację szerokiego spektrum cech dotyczących różnych obszarów funkcjonowania społeczeństwa przy maksymalnie wysokim pokryciu podmiotowym. Przygotowanie zbioru danych, który będzie wykorzystywał ogromny potencjał danych administracyjnych i umożliwiał tworzenie na jego podstawie nowych analiz i prowadzenie obserwacji zjawisk społeczno-ekonomicznych, wymaga zatem przeprowadzenia następujących etapów: przetwarzania, budowy operatu[15] oraz naliczenia wskaźników i danych wynikowych.

3.1.1. Przetwarzanie rejestrów administracyjnych

Proces przetwarzania rejestrów administracyjnych ma na celu podniesienie jakości danych z rejestrów urzędowych. W tym celu zostało stworzone w statystyce odrębne, zabezpieczone (posiadające kontrolę dostępu i szyfrowane), dedykowane środowisko informatyczne (Operacyjna Baza Mikrodanych). Jest to wydzielony obszar, w którym przetwarzane są dane pozyskiwane z rejestrów urzędowych, systemów informacyjnych administracji publicznej oraz niepublicznych systemów informacyjnych. Produktem końcowym procesu jest rejestr statystyczny, w którym przekształcone już dane są dostatecznie pełne pod względem podmiotowym oraz przedmiotowym i jednocześnie odpowiadają wprowadzonym na podstawie ustaw, standardom klasyfikacyjnym, nomenklaturom i definicjom podstawowych kategorii. Wyzwania jakie wiążą się ze standaryzacją i integracją danych zostały opisane w podrozdziale 3.1.

Dane pozyskiwane z rejestrów administracyjnych zostają poddane odpowiednim mechanizmom i procedurom zarówno informatycznym, jak też merytorycznym. W ramach procesu przekształcenia danych, standaryzacji i aktualizacji ulegają zmienne identyfikacyjno-adresowe oraz klasyfikacyjne. Jednocześnie, co bardzo istotne, ich wartość merytoryczna nie zostaje zmodyfikowana.

Każdorazowo, dane ze wszystkich źródeł administracyjnych pozyskiwanych przez statystykę publiczną zostają szczegółowo sprawdzone pod kątem występowania w nich zapisów powielonych, niedokładnych, nieaktualnych, błędnych bądź nieprawidłowo sformatowanych. Wszelkie nieprawidłowości poddane zostają procesom przekształcania składającym się z procedur i złożonych algorytmów: weryfikujących, korygujących i uzupełniających dane. Działania te wpływają bezpośrednio na jakość danych zwiększając ich poprawność, dokładność i przydatność, a także efektywność procesów integracji.

W pierwszym etapie procesu przetwarzania danych administracyjnych następuje weryfikacja zbiorów. Sprawdzane jest czy struktura zbioru i format danych odpowiadają ustaleniom

[15] Operat to wykaz wybranych, według określonych cech, osób prawnych, jednostek organizacyjnych niemających osobowości prawnej oraz osób fizycznych będących podmiotem obserwacji statystycznej wraz z ich identyfikacją adresową (Źródło definicji: ustawa z dnia 29 czerwca 1995 r. o statystyce publicznej (Dz. U. z 2021 r. poz. 955).

zawartym w umowie dotyczącej przekazywania danych (weryfikacja struktur pozyskanych zbiorów pod kątem ich zgodności z PBSSP).
Kontrola poprawności i kompletności oraz standaryzacja danych ma na celu podniesienie jakości danych wejściowych. Otrzymane zbiory sprawdzane są pod kątem występowania w nich zapisów powielonych (deduplikacja[16]), niedokładnych, nieaktualnych, błędnych oraz nieprawidłowo sformatowanych. Inne zadania podejmowane podczas procesu weryfikacji to m.in. kontrola kontekstowa (analiza spójności pól adresowych od poziomu województwa do poziomu ulicy względem rejestru TERYT); walidacja – poprawa wartości nieprawidłowych, zgodnie z algorytmami, (np. walidacja identyfikatorów takich jak PESEL, NIP, REGON wymaga sprawdzenia poprawności cyfry kontrolnej i wymaganej długości). Dane administracyjne, dopiero po obróbce za pomocą wyżej wymienionych procesów, mogą być wykorzystywane jako bezpośrednie źródło danych do badań, do imputacji, szacowania i uzupełniania danych w badaniu, do tworzenia i aktualizacji wykazów lub operatów do badań, do aktualizacji zmiennych statystycznych, a także kontroli jakości danych zebranych w badaniach w sposób standardowy.

W celu kontroli przeprowadzonych procesów w statystyce publicznej został zbudowany wewnętrzny System Ewidencji Zmiennych. Służy on do przechowywania informacji o zbiorach pozyskanych z rejestrów administracyjnych. Znajdują się w nim metadane opisujące administracyjne systemy informacyjne w układzie wielowymiarowym. Naliczane są statystyki jakościowe dla każdej zmiennej pozyskanej z rejestrów administracyjnych.
W systemie tworzone są raporty oparte na obliczonych statystykach jakościowych dla zmiennych i zbiorów danych, dzięki czemu możliwe jest badanie stopnia poprawy jakości uzyskanego podczas procesu przetwarzania danych. System składa się z dwóch głównych elementów: modułu obliczania danych w środowisku ich przetwarzania oraz modułu wyświetlania danych w środowisku Windows (za pomocą przeglądarki).

Opracowany i utrzymywany spójny model pozwala na szybkie i efektywne wykorzystanie szerokiego zakresu merytorycznego danych administracyjnych dostępnych cyklicznie oraz generowanie wyników zgodnych z oczekiwaniami odbiorców, a więc: rzetelnych, aktualnych, o dużej częstotliwości i na niskim (obecnie niedostępnym dla badań reprezentacyjnych) poziomie agregacji przestrzennej.

Bazując na danych wejściowych z niezależnych rejestrów administracyjnych, za pomocą nowoczesnych metod i algorytmów dotyczących przygotowania, przetwarzania, integracji oraz weryfikacji danych można w efekcie końcowym uzyskać spójne i rzetelne informacje statystyczne. Rozwiązanie to uwzględnia potrzeby użytkowników (odbiorców) i wpisuje się w model nowoczesnej statystyki, w której dane wynikowe powstają szybko, efektywnie i bez zbędnych opóźnień.

[16] Deduplikacja - proces usunięcia nadmiarowych danych na poziomie jednostki występującej w zbiorze.

Istotne dla uzyskania ciągłości prac i porównywalności wyników jest, aby te źródła danych były powtarzalne – zarówno jeśli chodzi o cykliczność ich generowania i przekazywania przez gestorów danych do organów statystyki publicznej, jak również pod względem przekazywanej struktury zbiorów danych i ich jakości.

3.1.2. Tworzenie operatów

W celu realizacji badań statystycznych konieczne jest stworzenie operatów do badań. Odbywa się to przez integrację wybranych źródeł administracyjnych. Głównym operatem, który jest kluczowym elementem jest operat osób. Do budowy operatu osób wykorzystuje się m.in. PESEL, Centralny Rejestr Podmiotów – Krajowej Ewidencji Podatników (KEP), Zakład Ubezpieczeń Społecznych (ZUS), Kasa Rolniczego Ubezpieczenia Społecznego (KRUS), system KRUSNAL, Narodowy Fundusz Zdrowia (NFZ) w zakresie osób ubezpieczonych; Agencja Restrukturyzacji i Modernizacji Rolnictwa (ARiMR), w zakresie osób fizycznych – producentów rolnych oraz posiadaczy zwierząt gospodarskich: świń, bydła, owiec i kóz.

Stosuje się również kilka kryteriów eliminacji osób niespełniających założeń definicyjnych populacji. Wyłącza się numery PESEL osób uznanych za zmarłe, zamieszkałe lub pobierające świadczenia poza Polską oraz występujące tylko w jednym rejestrze.
Następnie do populacji osób dołączane są zmienne adresowe, co wiąże się z koniecznością integracji zbiorów, a tym samym z wyborem metody łączenia. Jedną ze stosowanych metod jest metoda deterministyczna[17]. Kluczem łączenia pomiędzy poszczególnymi źródłami danych jest identyfikator PESEL, który ze względu na swą unikalność umożliwia łączenie zapisów z dwóch lub więcej źródeł poprzez jednoznaczne wskazanie rekordów dotyczących tych samych jednostek badania.
Oprócz głównego operatu możliwe jest również tworzenie odrębnych operatów dla subpopulacji. Mogą być to na przykład populacje obejmujące obszary: rodzin, aktywności ekonomicznej ludności, jak również operaty dotyczące innych podmiotowości: mieszkalnictwa czy przedsiębiorstw.

3.1.3. Obliczanie wskaźników i danych wynikowych

Zbudowanie operatów statystycznych uwzględniających różne podmiotowości umożliwia obliczenie wskaźników i danych wynikowych. Zakres podmiotowy zdefiniowany jest przez operaty, natomiast zakres przedmiotowy wiąże się z obszarem badawczym i obejmuje określoną tematykę będącą podstawą do obserwacji zjawisk społecznych, ekonomicznych i przestrzennych. Poszczególne obszary tematyczne zawierać mogą informacje o kompletnej, pełnej populacji (np. osoby mieszkające w Polsce) lub jedynie wąskiej subpopulacji (np. osoby z niepełnosprawnością). Generowane w ten sposób dane są wysokiej jakości, spójne i możliwe do udostępniania na niskich poziomach agregacji terytorialnej. Umożliwia to

[17] Metoda deterministyczna integracji danych polega na łączeniu danych za pomocą unikalnego klucza łączenia, np. poprzez numer PESEL lub NIP.

również tworzenie statystyk z wielu obszarów tematycznych, opisujących zróżnicowane zjawiska i opisywanie ich w sposób wielowymiarowy.

Wyniki niektórych realizowanych w statystyce publicznej prac wykorzystujących rejestry administracyjne udostępniane są jako statystyki eksperymentalne, dostępne na stronie internetowej GUS (Główny Urząd Statystyczny / Statystyki eksperymentalne).

3.2. Dobre praktyki w zakresie pozyskiwania i wykorzystywania danych administracyjnych przez statystykę publiczną

Statystyka publiczna ma zapisaną w ustawie rolę głównego koordynatora wszystkich systemów informacyjnych, natomiast gestorzy systemów różnie postrzegają tą rolę. Dobre praktyki w zakresie pozyskiwania i wykorzystywania danych administracyjnych przez statystykę publiczną obejmują przede wszystkim stałą współpracę z gestorem danych. Taka bezpośrednia współpraca pozwala na szybkie reagowanie na zmiany w zakresie przekazywanych informacji oraz na zmiany osób odpowiedzialnych za zbiór, ale też zapewnia wsparcie dla statystyki w analizie jakości przekazanych zbiorów.

Niezbędne są bardzo szczegółowe robocze ustalenia z gestorem co do zakresu nowych źródeł danych i uzgodnienia merytoryczne, w tym dokładne rozpoznanie zawartości podmiotowej i przedmiotowej otrzymanych zbiorów. Mimo to, w pierwszym etapie pozyskiwania nowego źródła danych może się zdarzyć, że pierwotnie wprowadzone zapisy w PBSSP będą wymagać zmian po przekazaniu pierwszych zbiorów. Ponadto zmiany mogą wynikać z tego, że zakres danych u gestora również ewoluuje (zwłaszcza w przypadku nowych źródeł). Ustalenie terminów przekazywania danych musi uwzględniać z jednej strony potrzeby statystyki (informacje wynikowe powinny być publikowane możliwie szybko), ale też fakt, że wcześniejsze otrzymywanie danych od gestorów może powodować, że zbiory są niekompletne lub gorszej jakości.

Istotne jest także zapewnienie bezpieczeństwa przekazywanych danych, w tym w momencie ich transferu pomiędzy instytucjami. Z tego względu, zwłaszcza w przypadku dużych zbiorów danych, kluczowa jest współpraca z informatykami po obu stronach. Przy przetwarzaniu danych pozyskanych ze źródeł administracyjnych niezbędne jest uwzględnienie w harmonogramach prac czasu potrzebnego na odpowiednie opracowanie otrzymanych danych, aby były gotowe do wykorzystania w badaniach statystycznych, co jest pracochłonne zwłaszcza przy zbiorach przekazywanych po raz pierwszy.

Na zakończenie procesu wykorzystania danych administracyjnych kluczowa jest odpowiednia komunikacja w przypadku regularnego publikowania wyników badań uzyskanych z przetwarzania źródeł administracyjnych. Istotne jest aby cały proces był maksymalnie transparentny dla użytkowników danych.

4. Formalne i prawne aspekty wykorzystania danych administracyjnych ze szczególnym uwzględnieniem potencjału Programu Otwartych Danych

4.1. Otwieranie danych publicznych w Polsce

W ostatnich latach swoje zaangażowanie na polu otwierania danych zintensyfikowały różne podmioty, w tym organizacje międzynarodowe, mobilizując instytucje publiczne do udostępniania gromadzonych zbiorów. W raportach tych organizacji zawarto m.in. ocenę poziomu otwarcia danych generowanych przez polską administrację publiczną oraz zalecenia na temat konieczności przygotowania kompleksowego dokumentu na rzecz otwierania danych w Polsce.

Na forum krajowym, w coraz szerszym zakresie, podejmowane są działania dotyczące otwierania danych publicznych. Ustawa z dnia 6 września 2001 r. o dostępie do informacji publicznej wskazała główne kierunki, które zoperacjonalizowane zostały w kolejnych aktach prawnych i dokumentach programowych. Zgodnie z tą ustawą każda informacja o sprawach publicznych stanowi informację publiczną, a dostęp do niej, z wyjątkami określonymi w ustawie, przysługuje każdemu użytkownikowi. Udostępnianie informacji publicznej (w tym dokumentów urzędowych) następuje w różnych trybach przewidzianych w ustawie, w tym m.in. w Centralnym Repozytorium Informacji Publicznej (CRIP) utworzonym na podstawie Rozporządzenia Ministra Administracji i Cyfryzacji z dnia 26 marca 2014 r. w sprawie zasobu informacyjnego przeznaczonego do udostępniania w Centralnym Repozytorium Informacji Publicznej[18]. W 2016 r. Rada Ministrów przyjęła uchwałę ustanawiającą Program otwierania danych publicznych (PODP), który stanowi m.in. wsparcie dla realizacji przepisów ustawy z dnia 25 lutego 2016 r. o ponownym wykorzystaniu informacji sektora publicznego.

PODP, koordynowany przez Ministra Cyfryzacji, kierowany jest do organów administracji rządowej oraz jednostek administracyjnych im podległych lub przez nie nadzorowanych. Szczegółowo reguluje on kwestie rozwoju prac zmierzających do poprawy jakości zbieranych i udostępnianych danych publicznych (niezależnie od miejsca ich przechowywania). Zgodnie z zapisami tego programu dane publiczne są rozumiane jako 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 kontekst lub interpretacja.

W celu spełnienia wymogów ww. programu, dane publiczne powinny być publikowane zgodnie z następującymi zasadami:

[18] Niniejsze rozporządzenie utraciło moc z dniem 17 czerwca 2017 r. na podstawie art. 39 ustawy z dnia 25 lutego 2016 r. o ponownym wykorzystywaniu informacji sektora publicznego (Dz. U. poz. 352) i zostało zastąpione Rozporządzeniem Ministra Cyfryzacji z dnia 23 sierpnia 2018 r. w sprawie zasobu informacyjnego przeznaczonego do udostępniania w centralnym repozytorium informacji publicznej (Dz.U. 2018 poz. 1790).

a) dostępności – możliwe do pobrania przez każdego bez żadnych ograniczeń i do dowolnych celów,
b) upublicznienia w wersji źródłowej – o maksymalnej szczegółowości, w oryginalnej i nie zmienionej postaci,
c) kompletności – udostępnione w całości,
d) aktualności – w możliwie najnowszej wersji i najkrótszym terminie po ich wytworzeniu,
e) przetwarzania maszynowego – w formatach i postaci możliwej do odczytu komputerowego,
f) upublicznienia w sposób niedyskryminujący – dostępne bez potrzeby rejestracji, weryfikacji tożsamości lub podpisywania umów,
g) braku ograniczeń prawnych – możliwe do wykorzystywania do dowolnych celów, bez konieczności ubiegania się o jakąkolwiek zgodę (nie mogą być to dane objęte prawami autorskimi, patentami, znakami towarowymi, tajemnicami),
h) niezastrzegalności - w formacie możliwym do powszechnego stosowania, bez kontroli i ograniczeń (licencyjnych) kreowanych przez określone podmioty.

Dla zintensyfikowania działań w obszarze otwartości danych, Rada Ministrów przyjęła Program Otwierania Danych na lata 2021-2027, krajową strategię obejmującą kluczowe zagadnienia dotyczące udostępniania i zarządzania danymi[19].

Dane statystyczne stanowią szczególny typ danych publicznych. W przypadku danych pochodzących z systemu statystyki publicznej, pierwszorzędne znaczenie ma kwestia tajemnicy statystycznej wynikającej z ustawy z dnia 29 czerwca 1995 r. o statystyce publicznej (Dz. U. z 2022 r. poz. 459). Z tego też względu niektóre spośród wyżej wymienionych zasad nie mają zastosowania do danych statystycznych.

Z otwartych danych można korzystać w dwojaki sposób. Po pierwsze z gotowych narzędzi przygotowanych przez podmioty publiczne dostępnych najczęściej na portalach publikujących otwarte dane np. portal dane.gov.pl (danepubliczne.gov.pl).

Organy administracji rządowej oraz jednostki administracyjne im podległe są zobowiązane do publikowania tam generowanych przez siebie danych publicznych. Na portalu tym znaleźć można również zasoby udostępniane przez jednostki samorządu terytorialnego, a także odnośniki do systemów lub aplikacji zawierających dane o charakterze publicznym lub aplikacji wykorzystujących te dane. Drugim sposobem pozyskiwania danych z publicznych zbiorów jest wykorzystanie np. interfejsów programistycznych aplikacji (API).

[19] 7 września 2021 r. opublikowano Ustawę o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego. Zastąpiła ona Ustawę z dnia 25 lutego 2016 r. o ponownym wykorzystywaniu informacji sektora publicznego. Nowa regulacja uwzględnia zmiany wynikające z Dyrektywy 2019/1024/UE z dnia 20 czerwca 2019 ws. otwartych danych i ponownego wykorzystania ISP.

Otwieranie danych niesie ze sobą wiele korzyści, takich jak:

- usprawnienie wymiany wiedzy i informacji pomiędzy poszczególnymi instytucjami poprzez nowe drogi dostępu do danych;
- wzrost ponownego wykorzystania zasobów informacji publicznych zarówno przez inne podmioty publiczne, jak i przez sektor prywatny;
- podniesienie poziomu innowacyjności poprzez tworzenie przez kreatywnych odbiorców zasobów danych (przedsiębiorcy, start-upy) nowych produktów i usług dla obywateli;
- poprawa wizerunku instytucji publicznych, wzrost zaufania społecznego oraz wspieranie budowy społeczeństwa obywatelskiego.

Założeniem PODP jest poprawa jakości i dostępności tych danych, a kluczowe w tym zakresie jest zobowiązanie do współpracy wszystkich instytucji administracji centralnej. Takim miejscem gdzie naturalnie mogłyby przenikać się ścieżki danych publicznych jest GUS. To instytucja predystynowana do zajmowania się problematyką danych, pozyskiwania z nich wiedzy, tak by w sposób uświadomiony kształtować rozwój społeczeństwa informacyjnego. Statystyka powinna nieść odpowiedzialność za budowę infrastruktury informacyjnej państwa i być koordynatorem tworzenia, zarządzania i rozbudowy ekosystemu danych. Istotnym jest posiadanie wiedzy, w jaki sposób projektowane są poszczególne systemy. Pozwala to na zabezpieczenie przed niepotrzebnym równoległym generowaniem innych systemów oraz przed posiadaniem zduplikowanych informacji. Ustawa o statystyce publicznej zobowiązuje wszystkie publiczne organy administracyjne do przekazywania takich danych. Jednocześnie każdy organ administracji publicznej ma swoje akty legislacyjne, co wiąże się z utrudnionym dostępem do tego rodzaju zasobów dla statystyki. Z czasem jednak, wraz ze wzrostem zrozumienia dla otwierania danych oraz z intensyfikacją współpracy pomiędzy jednostkami administracji publicznej, coraz łatwej będzie statystyce publicznej uzyskiwać dostęp do danych administracyjnych.

4.2. Otwarte dane - definicja

Definicja „otwartości” powinna być bezpośrednio odnoszona do wiedzy, wspierania współpracy, w którą każdy może być zaangażowany i w ramach której wartością nadrzędną jest interoperacyjność (Open Knowledge Foundation, 2022). Za dane otwarte uznaje się dane aktualne, stale dostępne dla każdego, możliwe do ponownego wykorzystania (w tym także komercyjnego), dostępne bez barier (ograniczeń prawnych lub technicznych), udostępniane w sposób umożliwiający ich automatyczne przetwarzanie (odczyt komputerowy), zapisane w odpowiednim formacie wraz z metadanymi, w jak najmniej przetworzonej formie.

Poziom otwartości danych jest zwykle określany za pomocą skali opracowanej przez Sir Timothy’ego Bernersa-Lee (5 Star Data, 2022), która odnosi się do sposobów i formatów, w jakich dane te są udostępniane. Poziom zaawansowania oznaczany jest za pomocą gwiazdek (od 0 – brak otwartości do 5 – najwyższy poziom), zgodnie z poniższym zestawieniem:

☆☆☆☆☆ - dane w formatach uniemożliwiających przeszukiwanie (np. skan strony w pliku PDF);

★☆☆☆☆ - dane udostępnione bez określonej struktury (np. DOC, PDF możliwy do przeszukiwania);

★★☆☆☆ - dane ustrukturyzowane, ale zamknięty format (np. XLS);

★★★☆☆ - użyty został otwarty format (np. CSV, XML);

★★★★☆ - dane posiadają ujednolicone identyfikatory zasobów (URI), do których można linkować;

★★★★★ - dane tworzą powiązania do innych danych – (LOD).

Analizując dostępne technologie oraz potrzeby użytkowników korzystających z dużych zbiorów danych, za najbardziej pożądane rozwiązanie uważa się obecnie udostępnianie danych dzięki API – Programistycznym Interfejsom Aplikacji, które umożliwiają automatyczną komunikację systemów dzięki „połączeniom" zdefiniowanym na poziomie kodów źródłowych aplikacji informatycznych.

W statystyce wiąże się to z tym, że zestawienia danych winny być generowane bezpośrednio z systemów ich przetwarzania, ich publikacja odbywać się powinna w systemach udostępniania danych, najlepiej z API, a wybrane tablice publikacyjne winny zostać przetworzone do formatu CSV (lub innych otwartych formatów pliku, którego struktura umożliwia automatyczne, "maszynowe" przetwarzanie danych). Dodatkowo metadane powinny być ustandaryzowane, a dane udostępniane na możliwie najniższym stopniu agregacji w pełnych przekrojach i pełnych dostępnych szeregach czasowych. Ponadto dane oraz metadane powinny być dostępne w języku angielskim.

Z budową interfejsów programistycznych do aplikacji informatycznych związane jest zapewnianie interoperacyjności pomiędzy systemami/rejestrami. W praktyce oznacza to, że dane mogą być łatwo poddawane ponownemu wykorzystaniu np. w innych rejestrach, umożliwiając spójne współdziałanie różnych systemów (Gonzales Morales, Orrell, 2018). Takie podejście pozwala także na usprawnienie przebiegu procesów oraz racjonalizację kosztów funkcjonowania instytucji publicznych.

4.3. Działania zwiększające poziom otwartości danych w statystyce publicznej

Instytucje publiczne odpowiadają za generowanie i gromadzenie ogromnych zbiorów danych. Dzięki rozwojowi w zakresie telekomunikacji możliwe jest zapewnienie dostępu do tych danych, a także ich dalsze wykorzystywanie, które sprzyja postępowi gospodarczemu poprzez tworzenie nowych dóbr i usług. Inicjatywa otwierania danych dotyczy udostępniania ich w sposób umożliwiający jak najbardziej efektywne wykorzystanie przez użytkowników. Zmienia się więc tradycyjne podejście do podstawowej działalności krajowych urzędów statystycznych. W dzisiejszych czasach, erze szybkiego wzrostu zasobów informacji i danych,

najważniejszym zadaniem jest pozyskiwanie kluczowych informacji, ich przetwarzanie i udostępnianie w sposób otwarty i przystępy dla odbiorców danych statystycznych. Realizując swoją misję GUS jest aktywnym uczestnikiem krajowej i międzynarodowej społeczności działającej na rzecz otwartości danych. Jako priorytet uznawane jest wykorzystywanie technologii informacyjnych wspierających dostęp do danych przez użytkowników zewnętrznych oraz inicjatywy na rzecz współdziałania systemów i gromadzonych w nich danych, zarządzanych przez instytucje publiczne.
W 2019 r. Prezes GUS przyjął strategiczny manifest „5 O" na rzecz otwartości i transparentności w statystyce publicznej[20]. Manifest „5 O" to: *open data* (otwarte dane), *open access* (otwarty dostęp), *open algorithms* (otwarte algorytmy), *open source* (otwarte oprogramowanie), *open knowledge* (otwarty dostęp do zasobów nauki) czyli zasady jakimi mają kierować się służby statystyki publicznej w opracowywaniu informacji statystycznych.
Podejmowane działania realizowane na rzecz otwierania danych znajdują odzwierciedlenie w utrzymującej się wysokiej pozycji GUS w przeglądach Open Data Inventory (ODIN) organizowanych przez organizację Open Data Watch (ODW). Według najnowszego rankingu ODIN, oceniającego stopień dostępności i otwartości danych prezentowanych przez krajowe urzędy statystyczne, GUS awansował na 2 pozycję wśród 187 krajowych urzędów statystycznych na świecie. Wskaźnik ogółem wyniósł 85 pkt i był gorszy jedynie od wyniku uzyskanego przez Singapur.

Mając na uwadze potencjał i korzyści wynikające z otwierania danych publicznych, GUS podjął szereg inicjatyw na rzecz udostępniania danych statystycznych zgodnie z zasadami otwartości i obowiązującymi standardami, w tym doskonalenie dostępności informacji publikowanych w otwartych formatach oraz zwiększanie liczby systemów/produktów udostępniających dane poprzez API.

W latach 2018-2020 GUS włączył się w realizację projektu „Otwarte dane – dostęp, standard, edukacja". Liderem tego projektu, współfinansowanego ze środków POPC, jest Ministerstwo Cyfryzacji, a celem zadania było tworzenie rozwiązań systemowych na rzecz zwiększenia otwartości danych publicznych. Głównymi produktami projektu są: budowa API do Banku Danych Lokalnych oraz poszerzenie możliwości wykorzystania danych udostępnianych za pomocą API. Oficjalna premiera API do BDL miała miejsce 4 grudnia 2018 r. W ramach tych działań zbudowany został również pakiet w języku programowania R umożliwiający wykorzystanie funkcjonalności programu R do analiz danych udostępnianych przez API BDL.

W celu ułatwienia dostępu do systemów GUS udostępniających dane poprzez API, została utworzona specjalna witryna: https://api.stat.gov.pl. Użytkownicy mają dostęp poprzez API do rejestrów: REGON oraz TERYT oraz do BDL. Obserwuje się wysokie zainteresowanie gromadzonymi na ww. witrynie informacjami oraz danymi udostępnianymi za pomocą

[20] Manifest został ogłoszony przez Prezesa GUS w 2019 r. na Konferencji Metodologii Badań Statystycznych MET2019.

interfejsów programistycznych. W okresie od 1 stycznia 2020 r. do końca grudnia 2021 z API udostępnionych rejestrów i BDL skorzystało:

- REGON – ok. 5,1 mld zapytań (8055 zarejestrowanych użytkowników, liczba zapytań w poszczególnych dniach tygodnia kształtowała się w granicach od 5-11 mln);
- TERYT – ok. 54,5 mln zapytań (847 zarejestrowanych użytkowników; liczba zapytań w poszczególnych dniach tygodnia kształtowała się w granicach od 6 do 477 tys.).
- w przypadku API do BDL – ok. 80 mln zapytań (3699 zarejestrowanych użytkowników, liczba zapytań w poszczególnych dniach tygodnia kształtowała się w granicach od 600 tys. do 9 mln).

Zgodnie z założeniami, wszystkie produkty/systemy GUS o charakterze bazodanowym będą docelowo udostępniały informacje poprzez API. Ten cel rozwija kolejny projekt realizowany obecnie przez GUS pn. *„Otwarte dane plus”*. Głównym zadaniem jest budowa interfejsu programistycznego API do zmodernizowanych Dziedzinowych Baz Wiedzy. API DBW wzbogaca dotychczasowy katalog produktów wyposażonych w ten interfejs (API REGON, API TERYT, API BDL, API SDG, API STRATEG). API DBW zostało udostępnione w marcu 2022 r.

4.4. Wyzwania związane z otwieraniem danych publicznych dla potrzeb polityk publicznych (monitorowania i ewaluacji)

Rozwój idei otwierania danych idzie w kierunku zwiększania ich ilości, w tym ilości danych wartościowych. Zwiększanie zasobów ma być realizowane poprzez: poprawę interoperacyjności i jakości danych (czyli zwiększanie danych dostępnych przez API); wzrost wykorzystywania i wymiany otwartych danych między instytucjami oraz ponownego wykorzystywania zasobów danych naukowych. W całym procesie istotne jest również wzmocnienie współpracy z krajowymi i zagranicznymi interesariuszami danych, zwiększanie świadomości społeczeństwa na temat potencjału otwartych danych oraz poprawa umiejętności pracowników administracji publicznej w zakresie otwierania i zarządzania danymi.

Umiejętne wykorzystanie danych z systemów administracyjnych jest warunkiem optymalizacji polityk publicznych. Jak zostało pokazane, zasoby zgromadzone w rejestrach publicznych są bardzo bogate, jednak w niewielkim stopniu wykorzystuje się je jako wsparcie polityk opartych na dowodach. Pewnym rozwiązaniem byłoby stworzenie systemu w zakresie wykorzystania zasobów informacyjnych państwa jako wsparcia dla monitoringu i ewaluacji polityk publicznych lub ogólnie badań naukowych.

W celu efektywnego wykorzystania informacji pochodzących z rejestrów administracyjnych najważniejsze jest doprowadzenie do integracji zasobów. Kluczową kwestią w tym obszarze jest przygotowanie procedur prawnych i organizacyjnych umożliwiających podjęcie takiego działania. Jest to kierunek niezbędny do dalszego otwierania danych z systemów administracyjnych.

Jako narzędzie umożliwiające integrację danych i usprawnienie analizy zjawisk społeczno-ekonomicznych może być wykorzystany Krajowy Rejestr Urzędowy Podziału Terytorialnego Kraju (TERYT). Identyfikatory systemu TERYT stanowią standard identyfikacji terytorialnej dla organów prowadzących rejestry urzędowe i systemy informacyjne administracji publicznej. Dzięki ich zastosowaniu możliwa jest integracja danych gromadzonych w różnych systemach oraz prowadzenie analizy zjawisk społeczno-ekonomicznych w różnych przekrojach.

Kluczową kwestią w wykorzystywaniu do analiz informacji pochodzących z rejestrów publicznych jest bezpieczeństwo danych i ochrona prywatności osób objętych rejestrem. Zasoby te były dotychczas udostępniane jedynie w postaci prostych informacji zbiorczych w ramach statystyki publicznej. Udostępniane dane mają charakter zagregowany, co nie pozwala na prowadzenie analiz na poziomie indywidualnym, a jedynie na zbiorowościach. Ponadto prezentowane wyniki pochodzą z pojedynczych źródeł bez możliwości łączenia informacji z różnych rejestrów publicznych.

Niektóre z gromadzonych danych badawczych to dane wrażliwe lub chronione prawem autorskim. Przepisy dotyczące ochrony danych oraz ogólne rozporządzenie o ochronie danych (RODO) wymuszają otrzymanie od uczestników badania zgody na gromadzenie i przyszłe ponowne wykorzystanie danych przez innych badaczy. Uczestnicy muszą być poinformowani, w jaki sposób dane badawcze będą przechowywane i wykorzystywane w perspektywie długoterminowej oraz w jaki sposób zostanie zachowana poufność. Dane wrażliwe i poufne można zabezpieczyć, regulując lub ograniczając do nich dostęp i ich wykorzystanie. Kontrola dostępu powinna zawsze być proporcjonalna do rodzaju danych i poziomu poufności.

Otwieranie danych statystycznych przynosi wymierne korzyści zarówno użytkownikom informacji, jak i instytucjom udostępniającym dane. Wdrożenie zasad i rozwiązań sprzyjających otwartości danych statystycznych motywuje do poszukiwania i sukcesywnego wdrażania najlepszych rozwiązań w tym zakresie. Zwiększanie poziomu otwartości danych i rozwój interoperacyjności systemów informatycznych w ramach systemu statystyki publicznej oraz budowa API służy usprawnianiu sposobu wykorzystywania danych statystycznych, jak również doskonaleniu efektywności przebiegu procesu produkcji statystycznej. Dalsze prace dotyczące wsparcia na rzecz rozwoju i ewaluacji polityk publicznych mogłyby być kierowane na działania z zakresu utworzenia API dla każdego systemu administracyjnego w zakresie danych jawnych.

Identyfikuje się też liczne wyzwania związane z dalszym otwieraniem danych. Do tych podstawowych zaliczyć można m.in.: duże zapotrzebowanie na wgląd do zasobów danych, rosnące ekosystemy danych oraz liczba osób zainteresowana tymi zbiorami, złożoność struktur, nowe źródła danych a przestarzałe tryby dostępu, brak umiejętności i kompetencji do korzystania z tych zasobów.

Rozwój otwartych danych jest nieunikniony i niezbędny. Wykorzystanie jego pełnych możliwości powstrzymywane jest przez pewne cechy otoczenia, na które poszczególne instytucje działające na rzecz otwierania danych mają ograniczony wpływ, takie jak:

1. Izolacja systemów i brak jednorodnych standardów – dane zbierane przez jedną instytucję, oprócz stosowania przez nią własnych standardów, systemów i rozwiązań informatycznych, nie wychodzą poza tą instytucję, a zakresy zbieranych danych często się powtarzają;
2. Niedostateczna polityka informacyjna w zakresie otwartych danych oraz niska świadomość użyteczności otwartych danych, bez których wykorzystywania podejmowanie decyzji i tworzenie polityk publicznych przez samorządy i rząd nie będzie odzwierciedlać pełni możliwości decyzyjnych;
3. Ograniczenia prawne związane z łamaniem prywatności czy bezpieczeństwa (RODO) – zgodnie z ideą otwartych danych wszystkie dane, które nie naruszają ochrony danych osobowych i tajemnicy handlowej mogą być udostępniane, choć wymaga to spełnienia określonych warunków prawnych, finansowych, organizacyjnych, technicznych i kulturowych (Belkindas, Swanson, 2014, s. 109; Zuiderwijk, Janssen, Davis, 2014, s. 22).
4. Polityka – zasadnicza część spośród danych statystycznych pozyskiwanych przez jednostki służb statystyki publicznej w ramach PBSSP stanowi dane o wysokiej wartości, będące informacjami sektora publicznego, których ponowne wykorzystywanie wiąże się z istotnymi korzyściami dla społeczeństwa, środowiska i gospodarki.

Trwający od kilku lat proces otwierania danych publicznych dopiero zaczyna się krystalizować. Równolegle na poziomie międzynarodowym pracuje się obecnie nad otwarciem części zasobów, których właścicielami są podmioty prywatne. Rolę statystyki publicznej rozbudowuje się w kierunku instytucji – opiekuna ładu informacyjnego, czyli nie tylko jednostki odpowiedzialnej za zbieranie danych i produkcję na ich podstawie pewnych wskaźników, ale też troszczącej się o jakość i bezpieczeństwo całego ekosystemu danych.

Ze względu na szczególną rolę statystyki publicznej w systemie informacyjnym państwa, jak również dotychczasowe doświadczenia i skuteczne modele współpracy GUS oraz instytucji realizujących zadania np. związane z oceną efektywności wsparcia przedsiębiorstw, statystyka publiczna powinna pełnić kluczową rolę w tym procesie. Zlokalizowanie w jednym miejscu danych pochodzących z różnych rejestrów i systemów stanowiłoby kamień milowy w obszarze monitorowania oraz ewaluacji polityk i interwencji publicznych.

5. Wykaz stosowanych skrótów

API – Application Programming Interface (interfejs programowania aplikacji)

ARIMR – Agencja Restrukturyzacji i Modernizacji Rolnictwa

BDL – Bank Danych Lokalnych

CIT – podatek dochodowy od osób prawnych

CRIP – Centralnym Repozytorium Informacji Publicznej

DBW – Dziedzinowe Bazy Wiedzy

DZD – Dziedzinowe Zbiory Danych

EKPS – Europejski Kodeks Praktyk Statystycznych

KEP – Krajowej Ewidencji Podatników

KRUS – Kasa Rolniczego Ubezpieczenia Społecznego

LOD – Linked Open Data (połączone otwarte dane)

NFZ – Narodowy Fundusz Zdrowia

ODIN – Open Data Inventory

ODW – Open Data Watch

ONZ – Organizacja Narodów Zjednoczonych

PBSSP – Program Badań Statystycznych Statystyki Publicznej

POPC – Program Operacyjny Polska Cyfrowa

POS – Plan Opracowań Statystycznych

PI – Portal Informacyjny

POPC – Program Operacyjny Polska Cyfrowa

PODP – Program Otwierania Danych Publicznych

REGON – Rejestr Gospodarki Narodowe

RODO – Ogólne rozporządzenie o ochronie danych, inaczej rozporządzenie o ochronie danych osobowych

RSI – Repozytorium Standardów Informacyjnych

SEZ – System Ewidencji Zmiennych

SOS – System Operatów Statystycznych

SPRA – System Przetwarzania Rejestrów Administracyjnych

TERYT – Krajowy Rejestr Urzędowy Podziału Terytorialnego Kraju

URI – Uniform Resource Identifier (ujednolicony identyfikator zasobów)

6. Bibliografia

6.1. Publikacje, artykuły, monografie

1. Belkindas M.V., Swanson E.V., 2014, *International Support for Data Openness and Transparency*, Statistical Journal of the IAOS 2, Vol. 30, Issue 2.
2. Beręsewicz M., Szymkowiak M., 2015, *Big data w statystyce publicznej – nadzieje, osiągnięcia, wyzwania i zagrożenia*, Ekonometria 2 (48), Poznań, s. 9-21.
3. Card D., Chetty R., Feldstein M.S., Saez, E., 2010, *Expanding Access to Administrative Data for Research in the United States*. American Economic Association, Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas.
4. Crawford, K., & Schultz, J. (2014), *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, Boston College Law Review, 55, 93-128.
5. Connelly R. Playford C.J., Gayle V., Dibben C., 2016, *The role of administrative data in the big data revolution in social science research*, Social Science Research 59 (2016), 1-12.
6. de Goes, J., 2013, *"Big Data" Is Dead. What's Next?*, VB/Big Data.
7. Demystifying big data: A practical guide to transforming the business of government, *TechAmerica Foundation*, 2012, Washington (Dostęp: https://statswiki.unece.org/download /attachments/80053387/Demistyfying%20Big%20Data.pdf?version=1&modificationDate=1374223553898&api=v2)
8. Elias P., 2014, *Administrative data*, [w:] Dus A., Nelle D., Stock G., Wagner G. (red.), *Facing the Future: European Research Infrastructures for the Humanities and Social Sciences*, SCIVERO, Berlin, 47-48.
9. Einav L., Levin J.D., 2013, The Data Revolution and Economic Analysis. National Bureau of Economic Research.
10. Goerge R. M., Lee, B. J., 2001, *Matching and cleaning administrative data*, [w:] Citro C. F., Moffitt R. A., Van Ploeg M. (red.), *Studies of Higher Population: Data Collection and Research Issues*, National Academies Press, Washington D.C., 197-219.
11. Groen J.A., 2012, *Sources of error in survey and administrative data: the importance of reporting procedures*, J. Off. Stat, 28, 173.
12. Holmes M., 1996, *The multiple dimensions of information quality*, „Information SystemManagement", vol. 13, no. 2.
    Kitchin, R., 2014a, *Big Data, new epistemologies and paradigm shifts,* Big Data Soc. 1, 2053951714528481.

13. Kitchin R., 2014b, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*, Sage, London.
14. Laney D., 2001, *3D Data Management: Controlling Data Volume, Velocity and Variety*, META Group Research Note 6.
15. Mascott L. 2006, *Ensuring the quality information, „KM Review",* vol. 8, no. 6.
16. Mergel I., Rethemeyer R.K, K. Isett, 2016, *Big Data in Public Affairs, Public Administration Review* (Dostęp: https://www.researchgate.net/publication/307431462)
17. Nowakowski M., 2015, *Ocena wiarygodności informacji w serwisach internetowych*, Studia Informatica, Zeszyty Naukowe Uniwersytetu Szczecińskiego, 36/863, Szczecin, s. 103-107.
18. Schroeder R., Cowls, J., 2014. Big Data, Ethics, and the Social Implications of Knowledge Production.
19. Stawecki T., 2007, Dostępność publicznych baz danych - stan dzisiejszy i kierunki zmian, BIULETYN BANKOWY, nr 1.
20. *Pokolenie gniazdowników w Polsce*. 2020. Urząd Statystyczny w Warszawie, Warszawa
21. Wallgren A., Wallgren B., 2014, Register-Based Statistics: Statistical Methods for Administrative Data, Second Edition, John Wiley & Sons, Ltd.;
22. Por. J. Wieczorkowski, M. Dałek, Problem przeciążenia informacyjnego a integracja systemów informatycznych, w: Europejska przestrzeń komunikacji elektronicznej, t. 1, „Zeszyty Naukowe", nr 762, „Ekonomiczne Problemy Usług", nr 104, Uniwersytet Szczeciński, Szczecin 2013, s. 439–448.
23. Woollard, M., 2014. Administrative data: problems and benefits. A perspective from the United Kingdom. In: Dus¸a, A., Nelle, D., Stock, G., Wagner, G. (Eds.), Facing the Future: European Research Infrastructures for the Humanities and Social Sciences. SCIVERO, Berlin.
24. Zuiderwijk A., Janssen M., Davis C., 2014, Innovation with Open Data: Essential Elements of Open Data Ecosystems, *Information Polity: The International Journal of Government & Democracy in the Information Age*, Vol. 19, Issue 1/2.

6.2. Referaty z konferencji

1. Gonzales Morales L., Orrell T., Data interoperability: A practitioner's guide to joining up data in the development sector – Global Partnership for Sustainable Development Data, the Second UN World Data Forum w Dubaju, październik 2018.
2. Referaty z Międzynarodowej Konferencji Ewaluacyjnej „Nowa rzeczywistość. Wyzwania dla polityki spójności i ewaluacji w czasach postpandemicznych. 27-

28.05.2021 r. https://www.ewaluacja.gov.pl/strony/xiv-miedzynarodowa-konferencja-ewaluacyjna/materialy-z-konferencji/ (Dostęp w dniu 31.03.2022 r.).

6.3. Akty prawne

1. Dyrektywa Parlamentu Europejskiego i Rady (UE) z dnia 20 czerwca 2019 w sprawie otwartych danych i ponownego wykorzystania ISP (2019/1024/UE)
2. Ustawa z dnia 29 czerwca 1995 r. o statystyce publicznej (Dz. U. z 2022 r. poz. 459) .
3. Ustawa z dnia 6 września 2001 r. o dostępie do informacji publicznej (Dz U. 2001 r. Nr 112 poz. 1198).
4. Ustawa z dnia 17 lutego 2005 r. o informatyzacji działalności podmiotów realizujących zadania publiczne (Dz. U. z 2021 r. poz. 670).
5. Ustawa z dnia 11 sierpnia 2021 r. o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego (Dz. U. z 2021 r. 1641).
6. Ustawa z dnia 6 marca 2018 r. - Prawo przedsiębiorców (Dz. U z 2022 r. poz. 24).
7. Rozporządzenie Ministra Cyfryzacji z dnia 23 sierpnia 2018 r. w sprawie zasobu informacyjnego przeznaczonego do udostępniania w centralnym repozytorium informacji publicznej (Dz.U. 2018 poz. 1790)
8. Uchwała Rady Ministrów z dnia 20 września 2016 r. w sprawie ustanowienia "Programu otwierania danych publicznych" (Nr 107/2016)
9. Europejski Kodeks Praktyk Statystycznych dla krajowych organów statystycznych i Eurostatu (organu statystycznego UE) przyjęty przez Komitet ds. Europejskiego Systemu Statystycznego Eurostatu w dniu 16 listopada 2017 r.

6.4. Źródła internetowe

1. 5 Star Open Data, https://5stardata.info/en/, Dostęp: https://5stardata.info (Dostęp w dniu: 10.03.2022r.);
2. Data Act, digital-strategyec.ueorpa.eu, Dostęp: https://digital-strategy.ec.europa.eu/en/ policies/data-act (Dostęp w dniu: 10.03.2022r.)
3. Open definition - Defining open in open data, open content and open knowledge, Open Knowledge Foundaton, Dostęp: https://opendefinition.org/od/2.1/en/ (Dostęp w dniu: 10.03.2022r.)
4. Repozytorium Danych Informacyjnych, rsi.gov.pl, Dostęp: https://rsi.stat.gov.pl (Dostęp w dniu 10.03.2022r.);
5. *Wykorzystanie źródeł administracyjnych*, Stat: Gov.pl, Dostęp: https://stat.gov.pl/badania-statystyczne/przekazywanie-danych-z-systemow/wykorzystanie-rejestrow-urzedowych-i-systemow-informacyjnych-administracji-publicznej-w-statystyce-publicznej/ (Dostęp w dniu 09.02.2022 r.);

6. Wykorzystywanie danych administracyjnych do ewaluacji polityk publicznych – wnioski i rekomendacje, ibs.org.pl, Dostęp: https://ibs.org.pl/publications/wykorzystywanie-danych-administracyjnych-do-ewaluacji-polityk-publicznych-wnioski-i-rekomendacje-2/ (Dostęp w dniu 09.02.2022 r.);
7. *"What is Big Data?",* Villanova University: http://www.villanovau.com/resources/bi/what-is-big-data/#.U85_uPbD-mc (Dostęp w dniu 09.02.2022 r.);
8. Open Data Watch, Open Data Inventory (ODIN) – pozycja Polski w rankingu 2020 r. https://odin.opendatawatch.com/Report/countryProfileUpdated/POL?year=2020 (Dostęp w dniu 14.03.2022 r.).