

Jolanta SZUTKOWSKA

Metadane w pracy statystyka

Praca z metadanymi¹ to praca ciągła, która wymaga integracyjnego, interdyscyplinarnego sposobu myślenia. Tendencje, jakie występują w metadanych wyznaczają kierunki prac, które już trwają, rozpoczynają się lub będą inicjowane. Aby zobrazować w sposób syntetyczny charakter tych prac, dla potrzeb artykułu, informacja na temat metadanych została sporządzona według następującego układu: rola metadanych w społeczeństwie informacyjnym, model metadanych, metadane a poszczególne fazy badania statystycznego, standaryzacja w metadanych, spojrzenie na metadane z różnych perspektyw, metadane a dziedzinowe bazy danych, metadane a podwyższanie jakości informacji statystycznej.

Rola metadanych w społeczeństwie informacyjnym to przede wszystkim zapewnienie szybkiego dostępu do informacji. Metadane definiowane są jako informacje o danych. Wiąże się to z tworzeniem zwięzłego i usystematyzowanego zestawu informacji odsyłającej, który służy do efektywnego wyszukiwania danych potrzebnych użytkownikowi.

Efektywne wyszukiwanie informacji nie jest proste, ponieważ zależy od wielu uwarunkowań. Złe zaprojektowana i opisana tablica może doprowadzić do chaosu informacyjnego zamiast zapewniać ład informacyjny. Metadana musi być więc kompletna i zrozumiała dla użytkownika, aby stała się efektywną „mapą drogową” w świecie informacji.

Metadanymi są m.in. zmienne i ich definicje, klasyfikacje, bazy danych, metodologia badań, informacje o jakości czy listy kodów. Informacje te są rozproszone w różnych bazach i dokumentach (np. publiczna hurtownia danych, Słownik Pojęć, program badań statystycznych statystyki publicznej, plan opracowań statystycznych, plan wydawniczy itp.).

Szeroki zakres informacji referencyjnych można pogrupować w syntetyczny sposób na trzy podstawowe kategorie²:

- definicyjne,
- metodologiczne,
- dotyczące jakości.

¹ Artykuł został oparty na doświadczeniach autorki z udziału w szkoleniu zagranicznym w Voorburgu.

² Zaproponowana typologia metadanych jest propozycją autorki przygotowaną ze względu na potrzeby artykułu.

Pojawia się również kategoria metadanych systemowych, operacyjnych, które powinny gromadzić informacje techniczne związane z funkcjonalnością środowiska metadanych.

Metadane definicyjne zawierają definicje pojęć i mierniki. Tu znajduje się Słownik Pojęć, którego celem jest zgromadzenie wszystkich określeń używanych przez statystykę publiczną do opisu zjawisk społeczno-ekonomicznych. Umożliwia to jednoznaczne rozumienie pojęć przez użytkowników danych statystycznych.

Metadane metodologiczne zawierają zwięzłe opisy metod stosowanych w badaniu statystycznym, np.: opis losowania próby, opis metody zbierania danych, przetwarzania danych (w tym imputacji danych), estymacji danych, dokumentację metodologiczną (publikacje z opisem metodologicznym, zeszyty metodyczne) itp.

Metadane dotyczące jakości danych obejmują syntetyczny opis danych według wybranego komponentu jakości (użyteczności, dokładności, terminowości i punktualności, porównywalności czy spójności). Dostępność i przejrzystość danych jest często opisywana w sekcji metadanych, która dotyczy polityki rozpowszechniania danych. Tam też znajdują się informacje o polityce poufności danych i tajemnicy statystycznej.

Zapewnienie efektywnego funkcjonowania metadanych zależy od wielu warunkowań, a przede wszystkim od stworzenia nowego sposobu myślenia, który polegać ma na:

- tworzeniu zintegrowanego podejścia do projektowania badań (w przeciwieństwie do traktowania każdego z badań jako niezależnego systemu) — wspólne dane wejściowe, wymiana danych;
- skumulowaniu wiedzy na temat badań w metainformacji zamiast opierania się tylko na wiedzy doświadczonych pracowników;
- wykorzystywaniu wspólnej struktury tablic dla wielu zmiennych z różnych źródeł i wyeliminowaniu projektowania tablic wyłącznie dla danych z jednego badania czy w obrębie jednej dziedziny statystycznej.

W metodach wykorzystywanych w Europejskim Systemie Statystycznym (ESS) planuje się w ciągu następnej dekady odejście od traktowania każdej dziedziny statystycznej jako niezależnego systemu. Model³ „od dołu do góry”, w którym statystyka w różnych dziedzinach tworzona jest niezależnie od siebie, jest słabo przystosowany do spełniania wymogów polityki zintegrowanych zbiorów danych. Tworzy się więc system, który będzie oparty na kompleksowych i zintegrowanych metodach w procesach statystycznych (europejska systemowa metoda statystyczna). Takie podejście gwarantuje dostarczenie kompleksowej informacji o zjawiskach społeczno-ekonomicznych opartej na różnych źródłach

³ Komunikat Komisji do Parlamentu Europejskiego i Rady w sprawie metod tworzenia statystyki UE — wizja na kolejną dekadę, dokument Komisji COM (2009) 404 wersja ostateczna, Bruksela 10.08.2009 r.

statystycznych, innowacyjnych metodach zbierania danych. Zwiększa to efektywność procesów statystycznych pod względem redukcji obciążenia respondentów i eliminowania powtarzających się danych oraz zmniejszania kosztów. Modyfikacja metod produkcji statystycznej wymaga jednak dalszych prac rozwojowych, zwłaszcza w zakresie zapewnienia poufności danych oraz ustalania relacji krajowych systemów statystycznych z ESS. Zmiany te mają wpływ na budowę systemu metadanych.

Korzyści ze stosowania metadanych to:

- a) ułatwienie zarządzania zasobami danych;
- b) szybkie uzyskanie informacji o zbiorach danych dotyczących wybranej dziedziny;
- c) wyjaśnianie konfliktów między zmiennymi (np. poznanie przyczyn, dlaczego liczba pracujących jest inna, gdy źródłem informacji jest spis powszechny, system rachunków narodowych czy badania rynku pracy);
- d) eliminowanie dublujących się informacji i harmonizacja informacji;
- e) usprawnienie działań polegających na uzyskiwaniu i aktualizacji danych;
- f) poprawa jakości i użyteczności danych oraz zwiększenie liczby użytkowników.

Istotnego znaczenia nabiera hierarchizacja pojęć w metadanych, ustalenie związków między nimi oraz wyszukiwanie informacji. Staje się ono łatwe, gdy posługujemy się precyzyjnym językiem i wiemy, czego szukamy, np. szukając informacji o pracujących możemy uzyskać informacje o pracujących, dla których źródłem danych jest system rachunków narodowych (z uwzględnieniem m.in. szarej strefy i gospodarstw domowych), pochodzące z badań poprzez przedsiębiorstwa oraz z badań gospodarstw domowych. Użytkownik decyduje, która informacja jest najbardziej przydatna do jego potrzeb. Przy podejmowaniu decyzji użytkownik bierze pod uwagę aspekty jakościowe danych, czyli ich kompletność, dokładność, terminowość, porównywalność, spójność.

Efektywne metadane powinny być proste, użyteczne, interoperacyjne, zdolne do rozszerzenia i doskonalenia, odzwierciedlać najważniejsze aspekty zjawiska.

MODEL METADANYCH

Terminologie, które często się pojawiają przy modelach metadanych to: Neuchâtel, Dublin Core i norma ISO 11179.

Neuchâtel to nazwa standardowego modelu klasyfikacji, która pochodzi od nazwy miasta w Szwajcarii, gdzie grupa ekspertów pracowała nad wypracowaniem tego modelu. Klasyfikacje w modelu Neuchâtel są usystematyzowane m.in. według obiektów, jak: rodzina klasyfikacji, klasyfikacje, wersje klasyfikacji, tablice korespondujące, poziom klasyfikacji, pozycja klasyfikacji, indeks klasyfikacji. Model klasyfikacji zawiera identyfikator, tytuł, opis, kontekst, obiekt, zespół obiektów.

Dublin Core (Dublin Core Metadata Element Set) oznacza zestaw elementów opisu dokumentów dostępnych w sieci (repozytorium cyfrowe, repozytoria tekstowe, multiwyszukiwarki).

ISO 11179 to międzynarodowy standard metadanych (Metadata Registry Standard).

W każdym modelu zmiennych metadanych pojawiają się podstawowe terminy, takie jak:

- **jednostka statystyczna** — określa podstawową jednostkę obserwacji w badaniu, np. przedsiębiorstwo, jednostkę lokalną, głowę gospodarstwa domowego;
- **populacja** — operat badania;
- **zmienna** — opis zjawiska ze względu na różne przekroje, np. bezrobotni w podziale na miasta i wieś;
- **jednostka pomiaru** — jednostka miary, w jakiej wyrażona jest mierzona wielkość danej zmiennej, np.: zł, szt., %;
- **klasyfikacje** — zestaw standardów klasyfikacyjnych, np. klasyfikacja: rodzajów działalności (NACE), zawodów (ISCO), edukacji (ISCED), statusu zatrudnienia, według płci — klasyfikacje wykorzystywane przez Eurostat, dostępne na serwerze RAMON;
- **lista kodów** — charakteryzuje daną zmienną (ten typ metadanych występuje najczęściej przy transmisjach danych do Eurostatu);
- **statystyczny pomiar** — określa rodzaj zastosowanej miary, np. średnia, mediana, indeks.

Rekomendowane przez ekspertów zagranicznych wzorce metadanych to rozwiązania stosowane przez statystykę norweską, kanadyjską i nowozelandzką.

METADANE A POSZCZEGÓLNE FAZY BADANIA STATYSTYCZNEGO

Występuje ścisły związek metadanych z poszczególnymi fazami badania statystycznego. W fazie planowania badania mamy do czynienia z rozpoznaniem w zakresie zmiennych występujących na formularzu oraz zmiennych wynikowych, a także mierników. Informacje te zasilają Słownik Pojęć. W fazie zbierania i przetwarzania danych pojawiają się metadane, które dotyczą metod: zbierania danych, wyboru próby i uogólniania wyników w przypadku badań losowych, edycji danych, schematu losowania, postępowania z brakiem odpowiedzi (w tym metod imputacji i estymacji). W fazie analizy pojawiają się metadane związane np. z opisem polityki rewizji danych i ich jakością. W fazie rozpowszechniania pojawiają się informacje, które dotyczą dostępności i przejrzystości danych, polityki poufności danych, dostępu do mikrodanych, prac z hurtownią danych i projektowaniem tablic w różnych układach.

Należy zwrócić uwagę, że przy udostępnianiu metadanych kluczową rolę odgrywają metody wyszukiwania informacji np. według definicji pojęć, podmio-

tów, dziedzin, klasyfikacji, które powinny ułatwiać szybki dostęp i wyjaśnianie wątpliwości w pracy z metadanymi (konflikty między zmiennymi).

STANDARYZACJA W METADANYCH

Standaryzacja odgrywa istotną rolę w metadanych, ponieważ standard stanowi sposób na takie same zrozumienie danych przez użytkownika krajowego i zagranicznego oraz na prostą i szybką wymianę danych wraz z opisami.

Standaryzacja może dotyczyć:

- formatu elektronicznego danych, czyli słownika wspólnego metadanych MCV⁴;
- struktury zawartości, np. struktura dokumentu elektronicznego w ramach administracji elektronicznej;
- narzędzia, np. system DDS⁵ stosowany w GUS — narzędzie informatyczne do zarządzania dokumentami;
- terminologii, np. lista dziedzin statystycznych według Komisji Ekonomicznej ONZ (UNECE).

Inne przykłady standardów to: normy ISO⁶, formaty do wymiany i rozpowszechniania informacji: SDDS⁷, SDMX⁸, ESMS⁹, XBRL¹⁰, wymiana danych z zakresu sprawozdawczości finansowej, Gesmes-Genedi, CSV¹¹ (przechowywanie danych w plikach tekstowych), model metadanych Neuchâtel wykorzystywany przy opisie klasyfikacji.

W ramach rozwoju i zwiększenia efektywności przekazywania oraz udostępniania danych i metadanych statystycznych pomiędzy organizacjami międzynarodowymi i ich krajami członkowskimi opracowywane są standardy formatów danych i metadanych oraz przewodniki. Dzięki tak opracowanej standaryzacji zapewniona jest porównywalność danych opisujących te same zjawiska. Jednym ze standardów, który odgrywa coraz większe znaczenie i jest promowany przez organizacje międzynarodowe jest standard SDMX.

Model SDMX jest modelem ogólnoświatowym stworzonym przez kilka organizacji międzynarodowych, m.in.: Europejski Bank Centralny, Eurostat, Międzynarodowy Fundusz Walutowy, OECD, ONZ, Bank Światowy. To standard do wymiany danych i metadanych.

Ze standardem SDMX związany jest format ESMS. Jest to, w syntetycznym ujęciu, 21 podstawowych pojęć, które pozwalają na opis procesu i produktu statystycznego oraz jego jakości.

⁴ MCV — the Metadata Common Vocabulary.

⁵ DDS — Dispositif de Documentation Structuree.

⁶ ISO — International Organization for Standardization.

⁷ SDDS — Special Data Dissemination Standard.

⁸ SDMX — Statistical Data and Metadata Exchange.

⁹ ESMS — Euro SDMX Metadata Structure.

¹⁰ XBRL — Extensible Business Reporting Language.

¹¹ CSV — Comma Separated Values.

ZESTAWIENIE PODSTAWOWYCH INFORMACJI DLA METADANYCH REFERENCYJNYCH W STANDARDZIE SDMX

Wyszczególnienie	Zakres informacji
Informacje kontaktowe	nazwa instytucji, departamentu, adres pocztowy oraz e-mail, fax, telefon
Aktualizacja metadanych	data załadowania metadanych i ich aktualizacji
Statystyczna prezentacja	opis źródła danych — zakres podmiotowy oraz przedmiotowy badania (wykaz zmiennych oraz definicje zmiennych i wskaźników); stosowane klasyfikacje; przekroje sektorowe (np. sektory własności, sektory ekonomiczne); rodzaj jednostki statystycznej (np. przedsiębiorstwo, gospodarstwo domowe, mieszkanie); populacja statystyczna (zbiór elementów podlegających badaniu statystycznemu); poziomy jednostek terytorialnych; badane okresy; terminy badania według poszczególnych faz, w szczególności: zbierania danych, przetwarzania i rozpowszechniania. Jeżeli niektóre z informacji zostały podane już w innym punkcie, należy dokonać do nich odwołania
Jednostka miary	stosowane w badaniu jednostki miary (m.in. pasażer, mieszkanie)
Okres sprawozdawczy	okres, dla którego zbierane są dane
Upoważnienie do zbierania, przetwarzania i upowszechniania danych — akty prawne, porozumienia, podstawy prawne	regulacje prawne i inne uregulowania, które stanowią podstawę do zbierania, przetwarzania i upowszechniania danych oraz wymiany danych
Poufność danych — polityka poufności, postępowanie z danymi	uregulowania prawne, które mają wpływ na politykę poufności danych oraz metody wykorzystywane do zapewnienia poufności danych
Polityka udostępniania danych	zasady polityki udostępniania danych z uwzględnieniem kalendarza udostępniania danych, form udostępniania danych
Częstotliwość rozpowszechniania danych	podanie, z jaką częstotliwością udostępniane są wyniki badań. Jeżeli informacja została podana w innym punkcie, należy dokonać do niej odwołania
Formy rozpowszechniania danych	podanie sposobów rozpowszechniania danych, np. opracowania sygnałne, publikacje, bazy danych, dostęp do mikrodanych, inne formy udostępniania
Dostęp do dokumentacji metodologicznej i dotyczącej jakości danych	opis dokumentów o charakterze metodologicznym: formularz, opis metodologiczny badania, w tym metody wykorzystywane w badaniu oraz dokumenty opisujące jakość danych — raporty jakości
Zarządzanie jakością danych — działalność ukierunkowana na zapewnienie jakości i ocenę jakości	przedstawienie zasad polityki jakości opartych na standardach jakości ESS oraz instrukcje, wytyczne, przewodniki, a także ogólna ocena jakości badania, np. na podstawie raportu jakości
Przydatność danych — potrzeby użytkowników, kompletność informacji	określenie zasadniczych kategorii użytkowników i ich potrzeb oraz dokonanie opisowej oceny stopnia zaspokojenia tych potrzeb pod względem kompletności przekazywanych danych, np. w stosunku do wymagań ESS
Dokładność i rzetelność — ogólna ocena dokładności, błędy losowe i nielosowe	ocena dokładności i rzetelności danych na podstawie syntetycznej analizy błędów losowych (wariancja, średni błąd standardowy dla kluczowych zmiennych) i nielosowych w badaniu, a w szczególności błędów pomiaru: oceny kwestionariusza, metody zbierania danych, pracy statystyków, postawy respondentów. Ponadto, informacje o błędach: pokrycia, braku odpowiedzi, w przetwarzaniu wyników badania, założeń do badania

**ZESTAWIENIE PODSTAWOWYCH INFORMACJI DLA METADANYCH REFERENCYJNYCH
W STANDARDZIE SDMX (dok.)**

Wyszczególnienie	Zakres informacji
Terminowość i punktualność	ocena terminowości badania z uwzględnieniem przestrzegania harmonogramu badania, a w szczególności podanie czasu, kiedy dane wynikowe mogą być udostępnione użytkownikom, w odniesieniu do okresu referencyjnego. Przy ocenie punktualności — podanie, czy kalendarz udostępniania danych jest respektowany (przestrzeganie terminów publikacji)
Porównywalność — przestrzenna i w czasie	ocena porównywalności przestrzennej z punktu widzenia wykorzystywanej klasyfikacji podziału terytorialnego, np. klasyfikacji NUTS, która zapewnia porównywalność geograficzną. Ocena porównywalności w czasie — podanie długości porównywalnych szeregów danych oraz przyczyny przerw w porównywalnych szeregach danych, jeżeli występują (np. zmiana definicji, zmiana klasyfikacji)
Spójność między dziedzinami i w ramach tej samej dziedziny	ocena spójności między dziedzinami, ze zwróceniem uwagi na wyjaśnienie spójności (lub jej braku) między badaniami o spójnych zmiennych, a także ocena w ramach badania i między badaniami w ramach tej samej dziedziny w przypadku występowania zmiennych spójnych
Koszt i obciążenie	podanie obciążenia w minutach czasu przepracowanego na przygotowanie i wypełnienie danych oraz koszt prowadzenia badania. W przypadku braku danych można podać liczbę podmiotów uczestniczących w badaniu
Rewizja danych — zasady polityki rewizyjnej oraz rewizja danych w praktyce	podanie, czy dane z badania są poddawane rewizji, jak często i z czego to wynika
Proces statystyczny	opis: rodzaju źródła danych; częstotliwości zbierania danych; metod zbierania i przetwarzania danych, ze szczególnym uwzględnieniem schematu losowania; kalkulacji współczynników uogólniających; błędów losowych w badaniach reprezentacyjnych oraz błędów nielosowych, metod walidacji i edycji danych (stosowanie założeń do kontroli logiczno-rachunkowej, imputacja danych, estymacja itp.); metod opracowania danych — agregaty; metod dostosowania danych (np. metody wyrównań sezonowych). Jeżeli niektóre z informacji zawarte są w innych punktach standardu SDMX, należy dokonać do nich odniesienia
Komentarz — uwagi, aneks	należy dokonać opisu tych aspektów badania, na które raz jeszcze chcemy zwrócić naszą uwagę, np. brak porównywalności danych z danymi z innych spójnych badań i podanie przyczyn, z czego to wynika, wskazanie metadanych, które mogą wykazywać związki z opisanymi metadanymi oraz dodatkowych dokumentów związanych z badaniem, które chcemy umieścić w aneksie

Źródło: opracowanie własne na podstawie opisu metadanych referencyjnych z badań udostępnionych na stronie Eurostatu: http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/metadata/metadata_link.

Opis metadanych według standardu SDMX może być wykorzystywany przez autorów badań do opisu konkretnego badania i udostępniany użytkownikom zewnętrznym. Stąd wynika jego znaczenie w udostępnianiu informacji o danych z publicznych baz danych (w tym dziedzinowych baz danych). Zakres informa-

cyjny wymagany do opisu badań statystycznych w standardzie SDMX wyznacza zakres niezbędnych metadanych, który powinien być zapewniony w środowisku metadanych w postaci opisów lub łączy do konkretnych baz, opracowań, publikacji, dokumentów programowych.

Korzyści ze stosowania standardów w metadanych są następujące:

- zmniejszenie obciążenia wymianą informacji, np. dane w tym samym standardzie są przekazywane równocześnie do Eurostatu i Międzynarodowego Funduszu Walutowego,
- zapewnienie spójności danych,
- zwiększenie efektywności przetwarzania danych,
- ograniczenie kosztów: narzędzia są bezpłatne, chociaż muszą być poniesione pewne koszty związane z wdrożeniem systemu (zdobycie niezbędnej wiedzy, adaptacja struktury IT),
- narzędzia są utrzymywane i systematycznie udoskonalone przez organizacje międzynarodowe,
- istnieje możliwość wdrożenia SDMX w ramach pomocy finansowej Unii Europejskiej (UE) (grant).

Należy zauważyć, że odpowiednia architektura systemu metadanych, która zapewnia funkcjonowanie różnorodnych informacji o danych w spójnym, zintegrowanym środowisku jest podstawą do prowadzenia dalszych prac rozwojowych z tego zakresu.

UPROSZCZONY SCHEMAT ELEMENTÓW METADANYCH STATYSTYCZNYCH

METADANE — DANE O DANYCH			
Operacyjna Hurtownia Danych	Słownik Pojęć	metadane definicyjne	standard SDMX
	program badań statystycznych statystyki publicznej	metadane metodologiczne	
	baza formularzy		
	zintegrowana baza obowiązków sprawozdawczych		
plan opracowań			
harmonogramy badań			
kalendarz udostępniania informacji wynikowych			
publikacje, zeszyty metodyczne			
analizy			
Publiczna Hurtownia Danych	opracowania z portalu edukacyjnego		
	ogólne informacje o wdrażaniu standardów jakości ESS w statystyce publicznej (informacja udostępniona na portalu informacyjnym GUS http://www.stat.gov.pl/gus/5466_PLK_HTML.htm)	metadane dotyczące jakości ¹²	
	raporty kompletności z badań		
raporty jakości			
mierniki jakości			

Źródło: opracowanie własne.

¹² Informacje zawarte w harmonogramach badań oraz kalendarzach udostępniania informacji wynikowej są wykorzystywane przy opisie jakości — terminowość i punktualność.

SPOJRZENIE NA METADANE Z RÓŻNYCH PERSPEKTYW

Powstawanie systemu metadanych nakłada na statystyków zadania związane z produkowaniem nie tylko danych, ale również metadanych. Prawidłowe funkcjonowanie metadanych wymaga pracy ciągłej i zespołowej ekspertów z wielu dziedzin i o różnej specjalizacji.

Wskazane byłoby, aby autorzy badań wykonywali systematycznie prace mające na celu prawidłowe funkcjonowanie metadanych. Prace dotyczą:

- doskonalenia opisów metodologicznych danych i zastosowanych mierników;
- aktualizacji zakresu definicji pochodzących z badań (praca w ramach Słownika Pojęć);
- tworzenia synonimów pojęć oraz ustalania podobieństw i różnic między pojęciami zbliżonymi. Funkcja synonimów pojęć odgrywa ważną rolę przy wdrażaniu funkcjonalnego wyszukiwania różnorodnych zestawów informacji, które użytkownik kojarzy najczęściej z nazwami potocznymi, jeżeli nie zna nazw specjalistycznych;
- przygotowywania opisów sposobów rozpowszechniania danych z zachowaniem zasad poufności (należy zwrócić uwagę na powiązanie z różnymi bazami danych);
- ciągłego aktualizowania danych i metadanych;
- opracowywania i doskonalenia dokumentacji z zakresu metodologii i jakości dla potrzeb metadanych (publikacje metodologiczne, zeszyty metodyczne, raporty jakości, mierniki jakości);
- dokonywania opisów wybranych cech jakości badania dla potrzeb metadanych, w zakresie:
 - użyteczności danych — identyfikacja potrzeb podstawowych użytkowników oraz ocena kompletności danych w stosunku do potrzeb,
 - dokładności danych — podanie podstawowych mierników jakości dostępnych z badania, takich jak: wskaźnik braku odpowiedzi i błędu standardowego, stopa imputacji, wskaźnik nadmiernego pokrycia,
 - terminowości i punktualności — podanie odstępu czasu między okresem referencyjnym a dostępnością danych wynikowych oraz dokonanie oceny przestrzegania punktualności planu wydawniczego,
 - porównywalności — podanie przyczyny występującej przerwy w dostępie do danych i wyjaśnienie jej (np. zmiana definicji, klasyfikacji),
 - spójności danych — wyjaśnienie różnic między danymi odnoszącymi się do badania tych samych zjawisk;
- opisów polityki rewizyjnej w zakresie danych;
- charakterystyki źródeł danych (chodzi tu o zwrócenie uwagi na to, czy jest to źródło pierwotne/wtórne, czy badanie jest pełne/losowe, czy badanie prowadzone jest poprzez przedsiębiorstwa/gospodarstwa domowe czy też badanie oparte jest na rejestrach administracyjnych);
- metodologii badania — zwrócenie uwagi na metody walidacji danych oraz sposoby edycji, imputacji i uogólniania wyników, a także sprecyzowanie, czy dane mają charakter surowy czy też podlegają wyrównaniu sezonowemu;

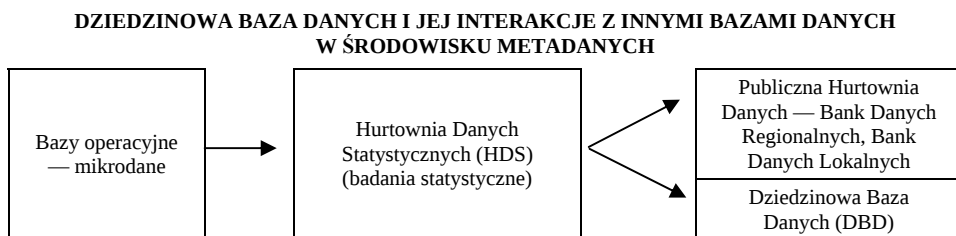
- kompletności metadanych pod względem merytorycznym (dotyczy to zapewnienia opisów tych aspektów badania, które są istotne dla jego specyfiki, np. opis polityki rewizji danych w systemie rachunków narodowych).

Do zadań metodologów należy nadzorowanie i przestrzeganie standardów oraz doskonalenie systemu metadanych we współpracy z autorami badań. Ponadto do ich zadań należy dbałość o jakość metadanych, czyli ich kompletność, integracja, eliminowanie konfliktów między danymi, wyeksponowanie różnic między danymi spójnymi, wyeliminowanie dublujących się informacji, transparentność metadanych.

Z kolei do zadań informatyków należy troska o sprawność serwerów oraz przyjazne środowisko metadanych, pozwalające na szybki dostęp do informacji, prosty sposób wyszukiwania, zastosowanie standardowego języka w metadanych, a także ich interoperacyjność. Natomiast służby udostępniania statystyki zapewniają doskonalenie dostępności i przejrzystości informacji, na podstawie wniosków płynących z informacji zwrotnej od użytkowników metadanych statystycznych.

METADANE A DZIEDZINOWE BAZY DANYCH

W kontekście prowadzonych rozważań nad metadanymi szczególne miejsce zajmują dziedzinowe bazy danych. Przy ich tworzeniu zalecane jest stosowanie podejścia interdyscyplinarnego. Należy zatem patrzeć na dziedzinę statystyczną z punktu widzenia jej powiązań wertykalnych i horyzontalnych z innymi dziedzinami i poddziedzinami.



Źródło: opracowanie własne.

Przy określeniu zakresu przedmiotowego ważne są zatem:

- inwentaryzacja źródeł danych,
- określenie układu hierarchicznego poddziedzin w ramach danej dziedziny,
- ustalenie związków dziedzinowej bazy danych z innymi dziedzinowymi bazami danych,
- zidentyfikowanie relacji dziedzinowej bazy danych z administracyjnymi źródłami danych.

Zakres przedmiotowy powinien zawierać zmienne i wskaźniki wynikające ze źródeł danych wykorzystywanych w dziedzinowej bazie danych, z uwzględnienie-

niem trendów rozwojowych w danej dziedzinie. Szczególną rolę mogą odegrać tu wskaźniki złożone, które łączą informacje z różnych dziedzin i poddziedzin, np. wskaźniki społeczeństwa informacyjnego czy zrównoważonego rozwoju. Przy określeniu trendów rozwojowych ważną rolę pełni analiza prac prowadzonych przez organizacje międzynarodowe i wnioski wynikające z analizy prac metodologicznych statystyki publicznej prowadzonych w ramach danej dziedziny, ze zwróceniem uwagi na powiązania danej pracy metodologicznej z prowadzonymi badaniami.

Zakres podmiotowy dziedzinowej bazy danych powinien zostać oparty na różnych źródłach danych (badania poprzez przedsiębiorstwa, gospodarstwa domowe, źródła administracyjne, badania pełne oraz reprezentacyjne).

Związek dziedzinowych baz danych z metadanymi definicyjnymi będzie wynikał głównie ze Słownika Pojęć.

Istotnego znaczenia nabiera fakt, że praca nad dziedzinowymi bazami danych będzie miała wpływ na doskonalenie Słownika Pojęć. Powinna nastąpić poprawa w zakresie:

- kompletności definicji pojęć, a w szczególności: definicji ogólnych, statystycznych i normatywnych oraz opisów źródeł definicji, szczególnie w odniesieniu do aktów międzynarodowych;
- aktualizacji pojęć — wprowadzone zostaną nowe pojęcia w ramach danej dziedziny oraz pojęcia złożone, związane z wieloma dziedzinami;
- dopracowywania opisów hierarchii pojęć na podstawie pojęć nadrzędnych i podrzędnych. Należy zwrócić uwagę, że ten rodzaj informacji, zawarty w Słowniku Pojęć, powinien być szerzej interpretowany niż jest obecnie. Przyczyni się to do zwiększenia jego przydatności przy ustalaniu relacji w ramach dziedzinowych baz danych oraz między dziedzinowymi bazami danych;
- doskonalenia opisów metod liczenia wskaźników. Chodzi tu o zwrócenie szczególnej uwagi na podanie wskaźników wraz ze wzorami oraz wprowadzenie jednostki miary;
- uzupełnienia zapisów dotyczących klasyfikacji i grupowań pomocnych przy ustalaniu przekrojów zmiennych w dziedzinowych bazach danych;
- rozszerzenia zestawu pojęć o pojęcia międzydziedzinowe;
- wykorzystywania Słownika Pojęć jako jednego z podstawowych narzędzi przy wyjaśnianiu konfliktów w pracy z metadanymi podczas wyszukiwania informacji. Jest to szczególnie istotne, gdy zaistnieje konieczność wyjaśniania rozbieżności wynikających z występowania różnych danych liczbowych dla spójnych zmiennych (wyjaśnianie podobieństw i różnic między definicjami zmiennych i ich wskaźnikami).

Związek dziedzinowych baz danych z metadanymi metodologicznymi będzie związany z zapewnieniem opisów metodologicznych z uwzględnieniem metod wykorzystywanych w badaniach statystycznych przedstawianych w publikacjach, opracowaniach, zeszytach metodycznych.

Z kolei związek dziedzinowych baz danych z metadanymi dotyczącymi jakości będzie wynikał głównie z opisów jakości źródeł danych na podstawie standardowych komponentów jakości (np. opis porównywalności danych historycz-

nych w dziedzinowej bazie danych, wyjaśnianie przerw w seriach danych). Opisy takie docelowo powinny znajdować się w raportach jakości z badań statystycznych i tablicach z miernikami jakości. Istotną kwestią będzie rozstrzygnięcie, czy przerwy w ciągach liczb można zastąpić w miarę możliwości imputacją danych.

W efekcie, opracowanie dziedzinowych baz danych umożliwi użytkownikom wyszukiwanie informacji według różnych słów-kluczy, np.: pojęć, wskaźników, dziedzin, poddziedzin, badań, jednostek uczestniczących w badaniu, klasyfikacji. Bazy te powinny być powiązane ze stronami urzędów administracji publicznej i organizacji międzynarodowych. Należy mieć na uwadze, że aktualizacja informacji w metadanych powinna odbywać się w sposób ciągły.

Wskazane jest doskonalenie standardu dokumentowania metadanych w dziedzinowej bazie danych, oparte na standardzie SDMX, zalecanym przez Eurostat. Standard ten zapewnia podstawowy zakres informacji dla użytkownika zewnętrznego.

METADANE A PODWYŻSZANIE JAKOŚCI INFORMACJI STATYSTYCZNEJ

Powstaje pytanie, czy budowa metadanych stworzy nową jakość w zarządzaniu informacją statystyczną? Związki metadanych z jakością to przede wszystkim: działania usprawniające, jakie wymusza tworzenie metadanych, polegające na integrowaniu danych z różnych źródeł; rozumienie spójności danych; identyfikowanie podobieństw i różnic między podobnymi danymi z różnych źródeł; eliminowanie danych powtarzających się i wyjaśnianie konfliktów między danymi oraz ukierunkowanie na uzyskanie coraz większej satysfakcji użytkownika.

Jakość w metadanych oznacza także stosowanie efektywnych standardów w metadanych i ich aktualizację.

Należy zwrócić uwagę, że przygotowanie użytkownika do posługiwania się metadanymi, jego wykształcenie i doświadczenie zawodowe będą miały wpływ na rodzaj problemów, jakie wystąpią w pracy z metadanymi i w konsekwencji — na opinie o jakości statystyki publicznej. Z tego wynika potrzeba odpowiedniego przygotowania użytkowników do pracy z metadanymi, np. poprzez organizowanie szkoleń.

Najczęściej występujące dylematy w pracy z metadanymi wiążą się z:

- wyborem źródła dla danej informacji (np. stopa bezrobocia z badania aktywności ekonomicznej ludności (BAEL) czy stopa bezrobocia rejestrowanego; ludność ze spisu czy ludność z badania aktywności ekonomicznej ludności);
- brakiem informacji dla danego przekroju, np. przedziałów wieku czy dochodów, rodzaju działalności, roku;
- brakiem danych aktualnych porównywalnych z poprzednim okresem, np. problem zmiany definicji, klasyfikacji;
- niezrozumieniem pojęcia (np. przyrost naturalny utożsamiany przez użytkownika ze wskaźnikiem dzietności);

- trudnościami z doбором metadanych wynikającymi ze złożonych pytań, np.: o jakość życia (czy wystarczy dokonać odniesienia do poziomu produktu krajowego brutto lub przedstawić mierniki pomiaru dobrobytu społecznego czy też należy posłużyć się miernikami zrównoważonego rozwoju), jakość pracy czy poziom informatyzacji kraju;
 - pomyłkami w formułowaniu zapytań, np. gdy użytkownik zamiast skrótu GDP¹³ użyje skrótu GPG¹⁴;
 - podaniem tego samego miernika w różnym ujęciu, np. wskaźnik strukturalny *gender pay gap* przedstawiany jako udział wynagrodzeń godzinowych kobiet do wynagrodzeń godzinowych mężczyzn oraz miara względna tego udziału jako 100% (wielkość udziału);
 - wyjaśnianiem konfliktów między metadanymi, gdy podobne informacje powtarzają się, ale mają przypisane różne wartości liczbowe.
- Przy wyjaśnianiu problemów z metadanymi ważną rolę będą odgrywać:
- praca ciągła metodologów, autorów badań, informatyków i użytkowników w doskonaleniu systemu metadanych (dobór zmiennych i mierników, ustalenie powiązań z innymi metadanymi);
 - Słownik Pojęć — przy precyzowaniu nazw zmiennych i wskaźników za pomocą ich definicji i metod kalkulacji, w przypadku pomyłek przy wyszukiwaniu informacji;
 - metadane metodologiczne zawierające m.in. opis źródła danych, publikacje, opracowania metodologiczne, w tym zeszyty metodologiczne przy rozpoznaniu, czy rzeczywiście te same zmienne się powtarzają czy też istnieją istotne różnice metodologiczne między zmiennymi, które „na pierwszy rzut oka” wydawały się identyczne;
 - metadane dotyczące jakości, zawierające m.in. informacje o porównywalności zmiennych i ich spójności (analiza spójności pod kątem zakresu przedmiotowego i podmiotowego badania, częstotliwości, sposobu zbierania danych, przekrojów).

Omówione uwarunkowania należy brać pod uwagę przy projektowaniu budowy metadanych, które powinny być przyjazne i łatwe w dostępie i obsłudze dla użytkownika.

Kluczową cechą metadanych jest ich interoperacyjność z innymi systemami. W celu doskonalenia systemu metadanych powinna być stworzona sekcja najczęściej zadawanych pytań FAQ¹⁵ oraz analizowana informacja zwrotna od użytkownika.

Podsumowanie

Celem artykułu jest syntetyczny przegląd różnych aspektów, które łączą się z pracami nad metadanymi, a przede wszystkim propozycja nowego sposobu

¹³ GDP — Gross Domestic Product.

¹⁴ GPG — wskaźnik strukturalny Gender Pay Gap.

¹⁵ Frequently Asked Questions.

myślenia o metadanych w kontekście prac związanych z budową zintegrowanego systemu informacji statystycznej i tworzeniem dziedzinowych baz danych.

W GUS prowadzone są prace nad doskonaleniem systemu metadanych w ramach budowy Systemu Informacji Statystyki Publicznej (SISP) oraz zespołów zadaniowych związanych z tworzeniem dziedzinowych baz danych.

Ważne jest zatem, aby statystycy, którzy uczestniczą w tworzeniu systemu metadanych, w tym dziedzinowych baz danych, zwrócili uwagę na następujące aspekty wykonywanych prac:

- praca z metadanymi ma charakter ciągły, a więc nie jest to praca wykonywana „od... do...”;
- w pracy z metadanymi uczestniczą m.in.: autorzy badań, informatycy, metodolodzy, eksperci ds. udostępniania informacji;
- zalecany jest interdyscyplinarny sposób myślenia, czyli identyfikowanie powiązań w ramach dziedzinowych baz danych, między dziedzinowymi bazami danych oraz z innymi bazami danych;
- w ramach dziedzinowych baz danych powinno nastąpić rozpoznanie otoczenia informacyjnego związanego z dziedziną statystyczną (inventaryzacja źródeł danych), wyznaczenie poddziedzin oraz ich hierarchizacja, wskazane jest rozpoznanie trendów rozwojowych w zmiennych i wskaźnikach. Ułatwia to tworzenie wizji rozwoju metadanych w kontekście ich powiązań z innymi dziedzinami;
- w ramach dziedzin i poddziedzin należy stosować różne źródła danych oparte na badaniach poprzez: przedsiębiorstwa, gospodarstwa domowe, administracyjne źródła danych, rachunki satelitarne;
- rola Słownika Pojęć w dziedzinowych bazach danych będzie wiązała się z:
 - poprawą opisów definicji pojęć pod kątem ich poprawności i kompletności,
 - zwiększeniem częstotliwości aktualizacji oraz weryfikacji merytorycznej tych opisów, ze zwróceniem uwagi na relacje między pojęciami nadrzędnymi i podrzędnymi,
 - doskonaleniem opisów wskaźników, szczególnie z uwzględnieniem algorytmów naliczeń i jednostek miar,
 - określeniem relacji wskaźników pod kątem powiązań hierarchicznych i horyzontalnych w ramach tej samej dziedziny i między dziedzinami, a także zaznaczeniem podobieństw i różnic w tych samych wskaźnikach pochodzących z różnych źródeł statystycznych (np. stopa bezrobocia z BAEL i stopa bezrobocia rejestrowanego),
 - rozszerzeniem zakresu przedmiotowego Słownika o pojęcia nowe w ramach dziedziny wynikające z tendencji rozwojowych i integracji różnych źródeł informacji, pojęcia złożone związane z wieloma dziedzinami, pojęcia ogólnostatystyczne dotyczące metadanych (w tym standardu SDMX) i jakości;
- rosnącego znaczenia będzie nabierać kategoria pojęć nadrzędnych i podrzędnych występujących w Słowniku Pojęć, która powinna być pomocna do identyfikowania relacji w zakresie dziedzinowych baz danych (można również

- rozważyć wprowadzenie nowej kategorii — pojęcia powiązane w kontekście identyfikowania relacji w ramach dziedziny i między dziedzinami);
- definicje ze Słownika Pojęć będą odgrywać kluczową rolę w wyjaśnianiu i usuwaniu konfliktów między metadany;
 - uzupełnieniu powinny ulec przekroje danych, które wiążą się ze stosowaniem klasyfikacji poziomych i hierarchicznych;
 - zdefiniowanie relacji serwera klasyfikacji z dziedzinowymi bazami danych;
 - tworzenie metadanych z zastosowaniem standardu SDMX, który wymusza spojrzenie na dane z punktu widzenia definicyjnego, metodologicznego i jakościowego;
 - stworzenie funkcjonalnego wyszukiwania dla większych zestawów informacji;
 - pojawianie się pojęć nowych nie tylko w ramach danej dziedziny, ale również pojęć powiązanych z kilkoma dziedzinami i ich ujęcie w metadanych, np.: rachunki satelitarne nauki, rachunki zdrowia, rachunki ochrony środowiska, dobrobyt społeczny, jakość zatrudnienia;
 - zintegrowanie prace na rzecz rozwoju metadanych prowadzone w ramach SISP i Dziedzinowej Bazy Danych;
 - interoperacyjność metadanych z innymi metadanymi (w tym metadanymi administracyjnymi);
 - rola metadanych jako instrumentu zarządzania informacją statystyczną, który prowadzi do harmonizacji, wyjaśniania konfliktów między metadanymi oraz eliminowania dublujących się danych.

mgr Jolanta Szutkowska — GUS

LITERATURA

Komunikat Komisji do Parlamentu Europejskiego i Rady w sprawie metod tworzenia statystyki UE: wizja na kolejną dekadę (2009), Komisja Europejska, Bruksela, 10.08.2009 r.

SUMMARY

A synthetic review of different aspects of metadata studies as well as tendencies of these studies are purpose of the article. These studies are continuous and need the integration as well as interdisciplinary way of thinking. Information on metadata is presented as follows: the role of metadata in society, the model of metadata, metadata and stages of a statistical survey, standardisation in metadata, different views of metadata (metadata and domain databases, metadata and improving statistical information quality).

РЕЗЮМЕ

Целью статьи является синтетический анализ разных аспектов работы по метаданным, а также тенденций определяющих направления этой работы. В статье изображается характер работы по метаданным — она имеет постоянный характер и требует интеграционного, интердисциплинарного способа мышления.

Информации по теме метаданных были представлены в разрезе: роль метаданных в информационном обществе, модель метаданных, метаданные и отдельные этапы статистического обследования, стандартизация в метаданных, подход к метаданным из разных сторон — метаданные и отраслевые базы данных, метаданные и повышение качества статистической информации.