Aleksandra Witkowska

STATYSTYCZNY SYSTEM INFORMACJI DUVA

1. Wprowadzenie

W administracji publicznej istnieje ogromny potencjał informacyjny. Równocześnie rośnie zapotrzebowanie na aktualne informacje, przy pomocy których można by wytyczać dalsze kierunki działania, zarówno w sektorze publicznym, jak i prywatnym. Podejmowane decyzje są coraz bardziej złożone, wymagają łączenia zasobów informacji z różnych dziedzin. Następują szybkie przemiany w życiu społecznym i politycznym. Stąd rosnąca rola urzędów statystycznych, które zajmują się gromadzeniem i przetwarzaniem informacji. W tej sytuacji już w latach osiemdziesiątych statystycy niemieccy razem z informatykami i przedstawicielami nauk społecznych podjęli próbę opracowania zasad kompleksowego statystycznego systemu informacji. Bodźcem do podjęcia tych działań było wprowadzenie nowoczesnych technik tworzenia banków danych, racjonalizacja sposobów opracowywania wyników, a tym samym rozszerzenie możliwości zaspokajania zróżnicowanych zamówień klientów.

Okazało się jednak, że występuje wyraźna dysproporcja między zasobami informacji a ich realnym wykorzystaniem. Niemieccy statystycy wielokrotnie podkreślali, że jedną z przyczyn tej rozbieżności jest fakt, że statystykę traktuje się najczęściej jako producenta liczb. Tymczasem jest to podejście zbyt wąskie. Wszyscy byli zgodni co do tego, że zadaniem statystyki jest zawsze próba opisu rzeczywistości za pomocą niewielu liczb. Najczęściej jednak z liczbami tymi wiążą się pewne treści, których statystyka mogłaby również dostarczyć. W związku z tym należałoby stworzyć taki system, którego elementem składowym byłaby część opisowa, dzięki której istniałaby możliwość udzielenia odpowiedzi na wszelkie dodatkowe pytania, na każdym etapie procesu badania statystycznego.

Powstał więc problem tak zwanych metadanych, czyli danych o danych, który stał się dla teorii i praktyki statystycznych systemów informacji podstawowym zagadnieniem badawczym w początkowej fazie ich rozwoju. Impulsem do rozpoczęcia kompleksowych badań nad takim systemem był coraz częstszy wymóg krajowej i międzynarodowej porównywalności danych, a także nowe możliwości w zakresie techniki przetwarzania informacji i komunikowania się przy równoczesnym dążeniu do obniżki kosztów działania urzędów statystycznych. Chcąc sprostać tym zadaniom podjęto wiele zespołowych badań w ramach Zrzeszenia KOSIS[1]. Jednym ze wspólnie przygotowywanych projektów było opracowanie zasad przeprowadzenia i wykorzystania wyników spisu ludności w 1987 roku. Program nosił nazwę Datenverarbeitungs-Unterstützung der Volkszählungs-Auswertung, stad późniejsza nazwa systemu DUVA. Wyniki uzyskane w ramach tego projektu uznano za zachęcające i podjęto decyzję uogólnienia zastosowanych instrumentów w taki sposób, aby były one użyteczne również w innych dziedzinach. Decyzja o podjęciu prac nad statystycznym systemem informacji DUVA zapadła w 1989 roku, kiedy podpisano porozumienie pomiędzy Zrzeszeniem KOSIS a miastem Stuttgart w sprawie rozwoju tego systemu. Stuttgart przyjął na siebie obowiązek finansowania prac nad projektem. Został również zobowiązany do powołania grupy badaczy, której zadaniem jest do chwili obecnej organizowanie i nadzorowanie wszelkich prac związanych z DUVA. W pracach nad omawianym systemem biorą udział statystycy z wielu ośrodków, przy czym znaczny udział, zwłaszcza na odcinku przetwarzania danych, ma Krajowy Urząd Statystyczny Berlina.

2. Cele i podstawowe elementy składowe systemu

Autorzy systemu DUVA wyszli z założenia, że zadania udostępniania danych, ich przygotowania, oceny i przekazywania określonym odbiorcom są, generalnie biorąc, jednakowe, bez względu na dziedzinę, której dotyczą. Gdyby zadania te oprzeć na jednakowych zasadach, na standaryzacji koncepcji i metod, na ujednoliceniu oprogramowania, to można by zmniejszyć czas i koszty potrzebne na realizację tych prac, a także łączyć informacje z wielu dziedzin.

[1] Zrzeszenie KOSIS jest komunalną organizacją samopomocy, która działa pod patronatem Związku Niemieckich Statystyków Miejskich (Verband Deutscher Staedtestatistiker) i wspierane jest przez Deutscher Staedtestag w zakresie wspólnych projektów dotyczących nowych technologii przetwarzania danych w dziedzinie statystyki i planowania. Do Zrzeszenia KOSIS należy ponad 100 okręgów administracyjnych i instytucji komunalnych, jak również wiele urzędów statystycznych.

Po długich dyskusjach sformułowano cele, jakie tworzony system ma spełniać i ustalono ich hierarchię.

Za podstawowy cel statystycznego systemu informacji uznano możliwie najlepsze zaspokojenie potrzeb informacyjnych płynących z parlamentu, administracji, gospodarki, nauki i opinii publicznej. Realizacja tego celu wymaga zorientowania całego systemu na potrzeby odbiorców. Wymóg orientacji na odbiorcę okazuje się jednak trudny w realizacji. Zapotrzebowanie na informacje statystyczne nie jest bowiem stałe, lecz zmienia się w miarę rozwoju gospodarczego i społecznego. Chcąc to zróżnicowane i zmienne zapotrzebowanie na dane w pełni realizować należy wszystkie dane, którymi dysponują Urzędy Statystyczne, potraktować jako wyjściowe bazy danych, uporządkować je i poddać standaryzacji. Tylko bowiem zbiór informacji elementarnych – mikrodanych – umożliwia realizację różnorodnych zamówień. Tak więc organizacja systemu mikrodanych jest drugim celem omawianego systemu.

Korzystanie z danych jednostkowych pociaga za sobą kolejny problem – zagadnienie tajemnicy statystycznej. Sformułowano więc kolejny cel, którym jest opracowanie zasad utajniania i dostępu do elementarnych danych, tak aby budowany system był zgodny z prawnymi zasadami funkcjonowania urzędów statystycznych, aby w pełni utrzymany został warunek pełnej anonimowości badanych podmiotów. W dalszej kolejności wymieniane są następujące cele:

– standaryzacja całego procesu produkcji statystycznej, której celem jest wspomaganie Urzędów Statystycznych, w tym racjonalizacja i przyspieszenie prac tych instytucji,

– doprowadzenie do pełnej integracji tego procesu na wszystkich etapach prac,

– rozwinięcie systemu metadanych i uzupełnianie go w miarę dodawania nowych elementów do programu,

– umożliwienie pracy użytkownikom w języku naturalnym,

– systematyczne rozpowszechnianie systemu i jego rozwój.

W związku z tym, że badanie statystyczne jest wieloetapowe zdecydowano, że również DUVA realizowana będzie w formie kolejnych elementów składowych – modułów. Moduły te mogą być stosowane niezależnie od siebie lub jako całość, w zależności od potrzeb indywidualnych odbiorców. Rozszerza to ewentualne grono użytkowników systemu.

Zasadniczym elementem w tym projekcie są modele danych (Datenmodelle). W ramach DUVA znajduje zastosowanie kilka takich modeli. Pierwszym z nich jest model danych podstawowych (Basisdatei). Jest to model utworzony dla danych pochodzących z administracji publicznej lub z badań bezpośrednich. Jakkolwiek zasady gromadzenia informacji w poszczególnych

miastach w ramach czynności administracyjnych (wydział meldunkowy, pomocy socjalnej czy komunikacyjny) czy też reguły stosowania określonych technik badawczych są takie same, gdyż opierają się na odpowiednich przepisach czy też przebiegają zgodnie z ustalonymi schematami, to jednak w praktyce zróżnicowany jest stopień dokładności pomiaru, klasyfikacje, aktualność zapisów czy stosowane jednostki miary.W związku z tym autorzy systemu sformułowali warunki, jakie muszą spełniać informacje z tego typu źródeł, aby można je wykorzystać jako bazy danych. Są one następujące:

– wszystkie informacje jednostkowe muszą pochodzić z tego samego źródła (wydział administarcyjny, instytucja czy też konkretne badanie statystyczne),

– muszą charakteryzować ten sam obiekt (osoba, mieszkanie, gospodarstwo domowe itp.),

– muszą być jednoznacznie określone pod względem czasowym i przestrzennym,

– wszystkie cechy muszą być zapamiętane każdorazowo w raporcie o stałych rozmiarach i w identycznym porządku,

– wszystkie informacje muszą być kompletne,

– dla wszystkich cech jakościowych muszą być opracowane wykazy klasyfikacyjne,

– wszystkie cechy jakościowe muszą być identycznie zakodowane,

– wszystkie cechy ilościowe muszą mieć jednoznacznie określone jednostki miary i ustaloną dokładność pomiaru,

– wszystkie informacje nowe, umieszczone po raz pierwszy w zbiorze, muszą być opisane w słowniku danych,

– wszystkie informacje dodatkowe, istotne do późniejszej interpretacji wyników, muszą być ujęte jako wolny tekst i zapamiętane oddzielnie.

Każda baza danych podstawowych może być przedmiotem agregacji, można w oparciu o nią tworzyć nowe cechy (jako kombinacje istniejących lub kombinacje wariantów cech), może stanowić element szeregów czasowych lub przestrzennych. Można też w niej dokonywać poprawek, jeżeli są konieczne, czy też poprzez odpowiednie grupowanie zmieniać nośnik informacji (mieszkanie – budynek). Równocześnie każda baza mikrodanych musi być odpowiednio opisana za pomocą określonych, umieszczonych w odpowiednim porządku metadanych. Szczegółową charakterystykę tego opisu przedstawili pracownicy Instytutu Demografii Stosowanej w Berlinie, Markus Guttler i Joachim Nguyen, w opracowaniu o „Das Statistische Informationssystem DUVA II" (Beschreibung, Analyse, Perspektiven).

W ramach systemu funkcjonować może wiele takich podstawowych baz danych. Stanowią one podstawę do budowy modelu makrodanych (Makrodatei). Zbiory makrodanych są w sensie technicznym podzbiorami należącymi

do zbiorów danych podstawowych. Pomyślane są jako etap pośredni pomiędzy danymi jednostkowymi a zestawieniami tabelarycznymi. Autorzy projektu uważają, że tradycyjny sposób zestawiania informacji w tabelach o różnych kombinacjach cech, dla różnych nośników informacji prowadzi do ogromnego nagromadzenia tabel, nie zawsze potrzebnych i dalej wykorzystywanych. Stąd też skonstruowali generator makrodanych, z wykorzystaniem którego można przygotować sobie bazę makrodanych zgodnie z indywidualnym zamówieniem. Podobnie jak zbiory danych podstawowych, tak i te podlegają określonym regułom standaryzacyjnym, między innymi:

– wszystkie cechy zbioru makrodanych są albo przejmowane bezpośrednio ze zbioru danych podstawowych, albo powstają przez ich grupowanie,

– wszystkie zbiory makrodanych są tworzone wyłącznie z jednego lub kilku zbiorów danych bazowych,

– zbiór makrodanych powstaje w ten sposób, że ustalona zostaje częstość występowania identycznych kombinacji cech w zbiorze podstawowym,

– konkretny zbiór makrodanych odnosi się do jednego typu nośników informacji. Może to być nośnik identyczny jak w zbiorze bazowym, ale może to też być nośnik o wyższym stopniu agregacji,

– wszystkie zbiory makrodanych, które będą przekazywane osobom trzecim do dalszego samodzielnego przetwarzania, muszą spełniać warunki określające zachowanie tajemnicy statystycznej.

Zbiory makrodanych stanowią, jak już zaznaczono, element pośredni, są podstawą do dalszej budowy tablic, wykresów i prowadzenia szczegółowych analiz statystycznych.

Kolejnym modelem danych w programie jest model określony jako tekst. Służy on do wprowadzenia opisów sytuacji prawnej, zasad zachowania tajemnicy statystycznej, wszelkich opisów stosowanych klasyfikacji, metod, definicji i informacji dodatkowych. W tym przypadku jako operatory używane są dostępne edytory tekstów.

Bardzo ważnym elementem składowym systemu jest słownik – Thesaurus. Składa się on z kilku części: listy używanych słów, listy odsyłaczy, katalogu i tablic kluczowych. Słownik skonstruowany jest tak, że można do niego dodawać nowe elementy, zmieniać istniejące, usuwać niepotrzebne lub uzupełniać opis. Czynności te wykonuje się wykorzystując dostępne edytory tekstów.

3. Praca z systemem

Zadaniem systemu DUVA jest, jak już wspomniano, wspieranie procesu produkcji statystycznej już od chwili planowania badania. Stąd też chcąc prze-

prowadzić konkretne badanie sprawdzamy najpierw w słowniku (Thesaurus) czy interesujące nas obiekty, cechy statystyczne są już tam wprowadzone i w jakim kontekście były analizowane. Jeżeli tego typu badanie nie było jeszcze prowadzone, to należy zapoznać się z warunkami, jakie musi spełniać baza danych podstawowych. Spełnienie ich gwarantuje nam możliwość dodania nowej bazy do systemu i dalszą pracę z wykorzystaniem wszystkich możliwości DUVA, a więc: generowanie makrodanych, zachowanie tajemnicy statystycznej, import danych do pakietów analizy statystycznej celem przeprowadzenia szczegółowych badań, budowa tablic czy wykresów, a także wnioskowanie z wykorzystaniem modelu – tekst.

Autorzy systemu podkreślają, że jego podstawowe zalety to praca w języku naturalnym, optymalizacja zasilania informacyjnego, racjonalizacja procesu produkcji statystycznej, zabezpieczenie tajemnicy statystycznej i otwartość na wszelkie nowości, zarówno od strony treści, jak i w zakresie nowych technik przetwarzania informacji.

DUVA różni się od innych systemów również tym, że nie dostarcza określonych, opracowanych wcześniej zestawień tabelarycznych czy wyników analiz, nie zawsze zgodnych z potrzebami klienta, ale w oparciu o dostępne bazy danych daje odpowiedź na pytanie: jakie obliczenia i oceny są w konkretnym przypadku możliwe do uzyskania. Dodatkowo, oprócz produktu końcowego, można uzyskać obszerne informacje o materiale wyjściowym, o warunkach dotyczących produkcji statystycznej (podstawy prawne, źródła danych, warunki zachowania tajemnicy statystycznej) czy też o innych zleceniach związanych z podobnym problemem badawczym.

4. Aktualny stan rozwoju systemu DUVA

Statystyczny system informacji DUVA jest już dosyć mocno rozwinięty. Wiele elementów składowych-modułów jest już dopracowanych i istnieje możliwość ich użytkowania. Do takich modułów należy moduł baz danych – Basismodul. Opracowana jest koncepcja baz danych podstawowych i makrodanych, a także korzystać można ze słownika. Jedynym mankamentem pracy w systemie w „naturalnym języku" jest to, że całość, zwłaszcza moduły: Thesaurus i Text opracowane są w języku niemieckim. Utrudnia to zastosowanie systemu w innych krajach. Autorzy DUVA liczą się jednak z możliwością przygotowania innych wersji językowych. Mają już dosyć zaawansowane prace nad opracowaniem wersji angielskiej. Przygotowany jest również generator makrodanych, i to w taki sposób, aby można go było stosować na pojedynczych stanowiskach oraz w sieci.

Kolejnym, gotowym już elementem systemu jest tzw. HERMES – program, którego zadaniem jest ułatwianie, czy też wspieranie prowadzonych badań (Erhebungsunterstützung). Za pomocą tego programu można ustalić i przechowywać adresy badanych jednostek statystycznych, uzgadniać terminy obserwacji, zapamiętywać cząstkowe wyniki obserwacji, a także ustalać stan zaawansowania danego badania w dowolnym momencie.

Mocno rozwiniętym fragmentem systemu, chociaż ciągle dyskutowanym i ulepszanym, jest program zachowania tajemnicy statystycznej. Nosi on nazwę SAFE.

W celu umożliwienia użytkownikom uzyskiwania z makrodanych produktów końcowych w postaci tabel, wykresów czy też wykonywania w oparciu o nie dalszych analiz opracowano zasady przejścia do generatora tabel DABANK-PC, a także do EXCEL-a, możliwe jest też wykonywanie dalszych analiz z wykorzystaniem pakietów SAS i SPSS.

Obok prac nad komponentami systemu zaawansowane są również prace nad koncepcją i szczegółową organizacją niektórych baz danych podstawowych. W pracach tych brali udział, obok członków zespołu roboczego DUVA, również pracownicy odpowiednich urzędów i instutucji odpowiedzialni za gromadzenie informacji. Przygotowano między innymi nowe kwestionariusze statystyczne, w których układ pytań, ich szczegółowość czy sposób kodowania informacji zostały na nowo zdefiniowane pod kątem potrzeb systemu. Do dyspozycji użytkowników są koncepcje następujących baz:

– ludność, jej stan i zachodzące zmiany – przygotowane dla działów ewidencji ludności,

– motoryzacja – dla wydziałów komunikacyjnych,

– budownictwo,

– poszkodowani – renciści,

– statystyka obcokrajowców,

– urzędowa pomoc socjalna – dla działów pomocy socjalnej przy urzędach miejskich,

– wypracowane są zasady zbierania i opracowywania danych spisów powszechnych (w oparciu o wyniki z roku 1987),

– aktualnie zakończona jest koncepcja gromadzenia danych o budynkach i mieszkaniach w związku z koniecznością przeprowadzenia takiego spisu w nowych krajach związkowych.

Wymienione wyżej koncepcje baz danych zostały już wdrożone w niektórych krajowych urzędach statystycznych. Koncepcje te, jak i wymienione wcześniej moduły systemu, mogą poszczególne instytucje czy urzędy zakupić, stając się tym samym użytkownikiem DUVA. Aktualnie pracuje z systemem lub wykorzystuje jego elementy ponad pięćdziesiąt okręgów administracyjnych. Rosnąca liczba użytkowników spowodowała konieczność wprowadzenia rejestru użytkowników. Rejestr taki, z pełną informacją o użytkownikach, będzie kolejnym elementem składowym tego systemu.