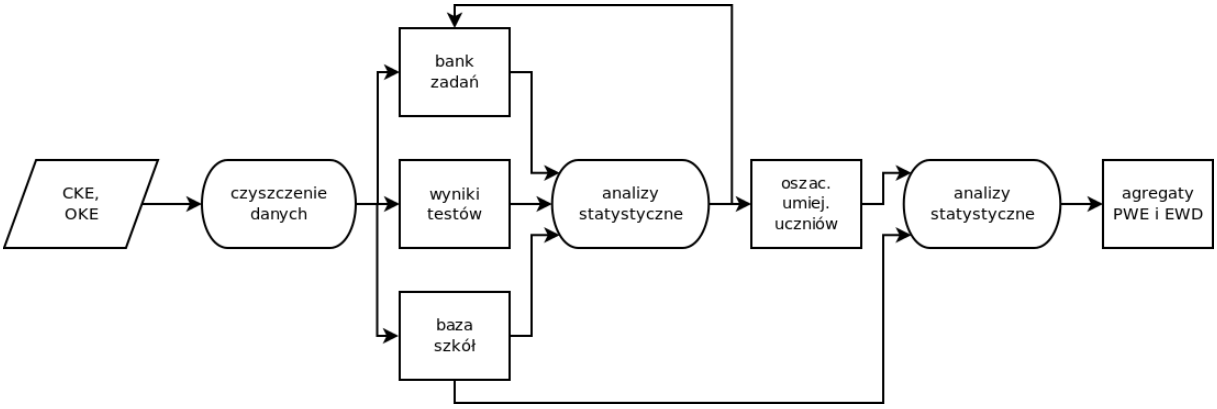
9. Baza danych IBE

W celu przeprowadzenia zrównywania, jak również w jego wyniku, zgromadzone zostały bardzo duża liczba danych, m.in. wyniki uczniów, oszacowania parametrów zadań i umiejętności uczniów, wskaźniki PWE dla szkół oraz jednostek samorządu terytorialnego itp. Aby z jednej strony zapewnić spójność wszystkich tych informacji, z drugiej zaś umożliwić łatwy i elastyczny dostęp do nich, stworzona została relacyjna baza danych oraz zestaw narzędzi ułatwiających korzystanie z niej. Ponieważ zakres gromadzonych i struktura wytwarzanych w ramach badania informacji w dużej mierze pokrywał się z projektem Edukacyjnej Wartości Dodanej (EWD), baza danych zaprojektowana została i wdrożona w taki sposób, aby objąć również dane zbierane i generowane w ramach projektu EWD. Co istotne, informacje znajdujące się w bazie danych są publicznie dostępne za pomocą kilku interfejsów dostosowanych do różnych grup odbiorców.

9.1. Dane zawarte w bazie danych

Dane zawarte w bazie podzielić można tematycznie na kilka grup: wyniki testów pisanych przez uczniów wraz z podstawowymi informacjami o uczniach, baza szkół, bank zadań, oszacowania umiejętności uczniów uzyskane w wyniku modelowania statystycznego oraz zagregowane wartości wskaźników PWE i EWD (Rysunek 9.1). Dokładny zakres dostępnych informacji omówiony został poniżej.



Rysunek 9.1 Schemat przepływu danych w bazie (grupy danych w bazie oznaczone prostokątami).

Podczas konstruowania bazy danych dużo wysiłku włożono w to, aby umożliwić łączanie danych pochodzących z różnych lat, a więc np. łączenie ze sobą wyników ucznia z egzaminów kończących kolejne etapy edukacji czy śledzenie zmian wyników w danej szkole, pomimo zmiany identyfikatora szkoły używanego w systemie egzaminów zewnętrznych[41]. W tym celu baza korzysta z własnych unikalnych identyfikatorów uczniów, szkół czy zadań, przy czym identyfikatory używane w innych systemach (np. kod OKE szkoły, identyfikator egzaminacyjny ucznia, itp.) także są w niej przechowywane i udostępniane.

[41] Np. wskutek zmian w strukturze administracyjnej kraju, zmiany organu prowadzącego, itp.

9.1.1. Wyniki uczniów i informacje o uczniach

W bazie przechowywane są wyniki wszystkich uczniów, którzy rozwiązywali arkusze standardowe następujących egzaminów:

- sprawdzianu szóstoklasisty z lat 2002-2014[42],
- egzaminu gimnazjalnego z lat 2002-2014[42],
- matury na poziomie podstawowym i rozszerzonym z biologii, chemii, fizyki i astronomii, geografii, fizyki, historii, informatyki, j. angielskiego, j. polskiego, matematyki oraz wiedzy o społeczeństwie z lat 2010-2014[42],
- matury poprawkowej z j. angielskiego, j. polskiego oraz matematyki z lat 2010-2014[42],

jak również wyniki testów przeprowadzonych w trakcie prowadzonych badań zrównujących.

Wyniki rejestrowane są na poziomie punktacji uzyskanej przez danego ucznia za każde z zadań, a w wypadku zadań ocenianych w podziale na kilka kryteriów (np. rozprawki z języka polskiego) punktacji uzyskanej za poszczególne kryteria. W odniesieniu do zadań zamkniętych jednokrotnego wyboru często dostępne są także informacje o tym, jaką dokładnie odpowiedź zaznaczył uczeń[43].

Poza samymi wynikami, zgromadzone zostały także dostępne w systemie egzaminów zewnętrznych dane kontekstowe o uczniach: płeć, rok urodzenia, posiadanie w trakcie danego egzaminu zaświadczenia o dysleksji oraz ew. informacja o byciu laureatem konkursu przedmiotowego[44].

Łączenie wyników uczniów pomiędzy latami i egzaminami możliwe jest:

- dla sprawdzianu od roku 2003,
- dla egzaminu gimnazjalnego od roku 2006,
- dla matury i matury poprawkowej od roku 2010.

9.1.2. Baza szkół

Zgromadzony w bazie wykaz szkół obejmuje:

- szkoły podstawowe z lat 2004-2014[42],
- gimnazja z lat 2004-2014[42],
- licea ogólnokształcące, technika oraz licea profilowane z lat 2010-2014[42].

[42] Do końca 2015 roku zostanie uzupełniona o wyniki z roku 2015.

[43] Dokładne informacje znaleźć można w tabeli na stronie http://zpd.ibe.edu.pl/doku.php?id=obazie.

[44] Co skutkuje przyznaniem uczniowi maksimum punktów za wszystkie zadania.

Dla każdej szkoły przechowywane są informacje o jej charakterystyce (szkoła dla młodzieży czy dla dorosłych, czy jest to placówka specjalna, czy jest to placówka związana[45)], historii zmian identyfikatora OKE szkoły, danych adresowych oraz przypisania położenia szkoły do gminy, powiatu i województwa w czasie, jak również liczbie mieszkańców miejscowości, w której znajduje się szkoła.

9.1.3. Bank zadań

Bank zadań przechowuje informacje o kryteriach oceny, schemacie punktowania[46], schemacie odpowiedzi (dla zadań jednokrotnego wyboru), występowaniu na testach (egzaminacyjnych i zrównujących) wraz z oznaczeniem poprawnej odpowiedzi w danym teście (dla zadań jednokrotnego wyboru), parametrach statystycznych (KTT oraz IRT), stosowanych w modelowaniu statystycznym przekształceniach skali (np. usuwaniu luk w punktacji, łączeniu kilku zadań w jedno, itp.) oraz, w większości wypadków, treści zadań (w formatach DOCX i HTML). Wyżej wymienione informacje dostępne są dla wszystkich zadań, które wystąpiły w arkuszach standardowych na jakimkolwiek z egzaminów, które wyniki znajdują się w bazie, jak również dla wszystkich zadań niepochodzących z egzaminów, które wykorzystane zostały w badaniach zrównujących (tzw. kotwica zewnętrzna).

9.1.4. Oszacowania umiejętności uczniów

Jednym z etapów zrównywania jest oszacowanie poziomu umiejętności uczniów na wspólnej skali za pomocą modelu IRT. Oszacowania takie są również wyliczane w procesie przygotowywania wskaźników EWD. Z racji znacznie lepszych od „surowych" wyników punktowych właściwości statystycznych tych oszacowań, polecane jest wykorzystywanie ich w miejsce sum punktów uzyskanych przez uczniów na egzaminach. Dostępne są oszacowania:

- estymatorami EAP oraz PV na zrównanej skali dla uczniów podchodzących do: sprawdzianu w latach 2002-2013, egzaminu gimnazjalnego w latach 2002-2013 oraz matury na poziomie podstawowym z j. angielskiego, j. polskiego i matematyki w latach 2010-2013.

- estymatorami EAP na oddzielnych skalach w każdym roku egzaminu dla: sprawdzianu z lat 2001–2011 oraz 2014[47], egzaminu gimnazjalnego z lat 2006-2014[42], jak również matury z j. polskiego, matematyki oraz łącznie wszystkich przedmiotów humanistycznych i wszystkich przedmiotów matematyczno-przyrodniczych z lat 2010-2014[42].

9.1.5. Zagregowane wartości wskaźników PWE i EWD

Przechowywane są wszystkie udostępniane w serwisach WWW (patrz podrozdział 9.2.1), jak również eksperymentalne i niepublikowane wartości zagregowanych wskaźników PWE i EWD. Wskaźniki PWE dostępne są dla szkół, gmin, powiatów, województw oraz Polski z lat 2002-2013 (dla sprawdzianu oraz egzaminu gimnazjalnego) lub 2010-2013 (dla matury na poziomie

[45] Typowym przykładem placówek związanych mogą być szkoły przyszpitalne.

[46] Także informacje o występowaniu luk w schemacie punktacji (np. język w ocenie wypracowania z j. polskiego na maturze punktowany na skali 0-1-3-6-9-12 punktów).

[47] Do końca 2015 roku zostanie uzupełniona o wyniki z roku 2012 oraz 2015.

podstawowym z j. angielskiego, j. polskiego i matematyki). Z kolei wskaźniki EWD obejmują okresy 2006-2014 dla gimnazjów[48] oraz 2010-2014 dla liceów ogólnokształcących i techników.

9.2. Dostęp do danych znajdujących się w bazie

Baza danych udostępniana jest na kilka sposobów (za pomocą kilku interfejsów). Różnią się one między sobą na wielu płaszczyznach: zakresem udostępnianych danych, łatwością użycia, możliwością (bądź jej brakiem) integracji z zewnętrznym oprogramowaniem. Zróżnicowanie to ma na celu dostosowanie do potrzeb bardzo zróżnicowanych grup odbiorców od osób zainteresowanych systemem edukacji (np. rodziców, nauczycieli i dyrektorów szkół, pracowników samorządów i administracji rządowej), poprzez naukowców, kończąc na programistach. Poniżej pokrótce omówiono każdy z nich.

9.2.1. Serwisy WWW

Serwisy http://pwe.ibe.edu.pl (omówiony w Rozdziale 8), http://ewd.edu.pl/gimnazjum, http://ewd.edu.pl/matura, umożliwiają samodzielne prowadzenie analiz na zagregowanych wskaźnikach Porównywalnych Wyników Egzaminacyjnych oraz Edukacyjnej Wartości Dodanej. Ich największa zaleta to prosty i intuicyjny sposób obsługi, który nie wymaga od odbiorcy umiejętności wykraczających ponad codzienne korzystanie ze stron internetowych. Dzięki temu są one dostępne dla wszystkich, także np. nauczycieli, rodziców czy przedstawicieli administracji, którzy nie posiadają zaawansowanych umiejętności statystycznych. Omawiane serwisy koncentrują się na analizie wartości wskaźników PWE i EWD za pomocą wykresów, natomiast dostęp do wizualizowanych danych, jakkolwiek możliwy, pełni rolę drugorzędną. Ich największa wada to ściśle ograniczony zakres udostępnianych danych – nie da się za ich pomocą pobrać z bazy danych nic więcej, niż zagregowane wartości wskaźników PWE i EWD.

Do tej kategorii zaliczyć należy również stronę WWW umożliwiającą przeglądanie banku zadań (w tym również pobieranie treści zadań czy podgląd ich parametrów statystycznych) znajdującą się pod adresem http://zpd.ibe.edu.pl/doku.php?id=bazatestypytania. Także w tym wypadku łatwość korzystania z serwisu okupiona jest ograniczeniem zakresu udostępnianych danych do ściśle określonego fragmentu bazy.

9.2.2. API HTTP

Kolejną metodą dostępu do danych zgromadzonych w bazie to API[49] HTTP. Jest to zestaw poleceń, dostępnych za pośrednictwem protokołu HTTP (tego samego, za pomocą którego przekazywana jest treść stron internetowych), które umożliwiają: przeszukiwanie banku szkół oraz informacji o podziale terytorialnym kraju, wyszukiwanie wskaźników PWE oraz EWD dla interesującego nas zakresu lat, egzaminu czy typu szkoły, pobieranie zbiorów danych z wartościami wskaźników PWE lub EWD dla zadanych szkół i/lub jednostek samorządu terytorialnego oraz wizualizację tych zbiorów danych w postaci wykresów (analogicznych do wykresów dostępnych w omawianych powyżej serwisach WWW). Interfejs ten powstał z myślą o programistach, którzy chcieliby w prosty i wygodny sposób zintegrować prezentację wskaźników PWE i/lub EWD z prowadzonymi przez siebie serwisami internetowymi. W oparciu o niego działają np. serwisy http://pwe.ibe.edu.pl,

[48] Do końca 2015 roku zostanie uzupełniona o okres 2013-2015.

[49] Application Programming Interface, pol. interfejs programistyczny.

http://ewd.edu.pl/gimnazjum oraz http://ewd.edu.pl/matura. Dokładny opis interfejsu znaleźć można na stronie http://zpd.ibe.edu.pl/doku.php?id=api_http.

9.2.3. Pakiet ZPD dla R

Poważnym ograniczeniem wymienionych powyżej sposobów dostępu do danych był stosunkowo wąski zakres możliwych do pobrania informacji. Pakiet ZPD dla R opracowany został z myślą o tym, aby udostępnić możliwie jak najszerszą część danych, jednocześnie nie wymagając od użytkownika nadmiernej wiedzy technicznej ani dokładnej znajomości fizycznej struktury bazy danych. Adresowany jest przede wszystkim do naukowców i analityków, ale także innych osób, dla których zakres danych i analiz udostępnianych w opisanych wyżej serwisach WWW jest niewystarczający.

Licząca blisko 100 tablic fizyczna struktura bazy danych została w pakiecie ZPD dla R uproszczona do 11 ułożonych tematycznie *grup danych*: wyniki testów (grupa *wyniki)*, baza uczniów (*uczniowie*), informacje o uczniach w kontekście konkretnego egzaminu (*uczniowieTesty*), zastosowane modele statystyczne (*skale*), oszacowania umiejętności uczniów (*oszacowania*), baza szkół (*szkoły*), baza testów (*testy*), bank zadań (*kryteriaOceny*), parametry statystyczne zadań (*parametry*) oraz zagregowane wartości wskaźników PWE i EWD (*wartościWskaźników*). Dane z poszczególnych grup dają się ze sobą łatwo łączyć dzięki występowaniu w nich wspólnych identyfikatorów (uczniów, zadań, szkół, itd.).

Istotną zaletą pakietu ZPD dla R jest dostępność funkcji automatyzujących najczęściej wykonywane czynności, np. obliczających sumę punktów z testu, normalizujących wyniki egzaminu (np. do skali *z* czy skali staninowej), odnajdujących dla każdego ucznia jego pierwsze lub ostatnie podejście do wskazanego egzaminu czy poprawnie agregujących wskaźniki PWE. Możliwość skorzystania z tych funkcji pozwala nie tylko przyspieszyć pracę z danymi, ale także ustrzec się błędów, jakie mogłyby się wkraść przy samodzielnym wykonywaniu tych, niekiedy dość złożonych, przekształceń.

Pakiet ZPD dla R został bardzo dobrze udokumentowany. Na stronie http://zpd.ibe.edu.pl/doku.php?id=r_zpd znaleźć można dokładne opisy poszczególnych *grup danych* i relacji między nimi wraz z przykładami użycia. Dostępny jest tam także wykaz wszystkich dostępnych zmiennych wraz ze wskazaniem, w jakich *grupach danych* występują oraz jakie dokładnie informacje przechowują. W końcu dostępne są tzw. samouczki, czyli rozbudowane przykłady na pobranie i przetworzenie danych znajdujących się w bazie do samodzielnego wykonania i przeanalizowania. Dokumentację uzupełnia omówienie bardziej zaawansowanych aspektów użycia pakietu ZPD dla R, takich jak samodzielne obliczanie zagregowanych wskaźników PWE czy dyskusję nad sposobami minimalizacji czasu pobierania danych z bazy.

Na koniec wypada wyjaśnić, dlaczego zdecydowano się na implementację tego interfejsu akurat w programie statystycznym R, a nie np. w bardziej popularnych w Polsce, jak SPSS, SAS albo Stata. Wymienić można dwie główne przyczyny. Pierwszą była chęć skorzystania z darmowego oprogramowania. Mamy nadzieję, że w ten sposób poszerza się grono potencjalnych użytkowników opisywanego interfejsu dostępu do bazy, nie każdego musi być bowiem stać na zakup licencji na komercyjne oprogramowanie statystyczne. Drugim czynnikiem była minimalizacja czasu niezbędnego do przygotowania interfejsu. Również to kryterium, dzięki doskonałej integracji z relacyjnymi bazami danych[50], R spełniał najlepiej.

[50] W szczególności dostępności pakietu dplyr.

9.2.4. Bezpośrednie wykonywanie zapytań SQL na bazie

Ostatnią możliwością dostępu do danych w bazie jest bezpośrednie połączenie z oprogramowaniem serwera bazy danych (używany system baz danych to PostgreSQL) i samodzielne formułowanie zapytań do bazy w języku SQL. Jakkolwiek daje to największe możliwości zarówno pod względem zakresu dostępnych danych (de facto cała baza), jak i potencjalnego zredukowania czasu ich pobrania, wymaga od użytkownika znacznej wiedzy z zakresu relacyjnych baz danych i biegłości w operowaniu językiem SQL. Nieumiejętne korzystanie z tego interfejsu rodzi niebezpieczeństwo pobrania innych danych, niż się zamierzało, jak również pobierania ich w taki sposób, który nie będzie się w stanie zakończyć w rozsądnym czasie. Dodatkowym utrudnieniem jest brak dokładnej dokumentacji. Dostępny jest co prawda diagram struktury fizycznej[51], na którym oznaczono wszystkie tablice wraz z kolumnami, relacje między tablicami, klucze podstawowe oraz indeksy, brak jednak opisów poszczególnych tablic i kolumn[52]. W zdecydowanej większości wypadków te same dane da się łatwiej i w krótszym czasie pobrać używając pakietu ZPD dla R i zalecane jest korzystanie właśnie z niego. Pomimo wszystkich wymienionych obostrzeń istnienie tego interfejsu jest niezbędne, jest on bowiem wykorzystywany wewnątrz pakietu ZPD dla R oraz API HTTP. Warto nadmienić, że za pośrednictwem ODBC można w ten sposób uzyskać dostęp do bazy z większości programów używanych do przetwarzania danych (zarówno programów statystycznych jak SPSS, Stata czy R, jak i np. z arkuszy kalkulacyjnych, choćby MS Excel).

[51] http://zpd.ibe.edu.pl/lib/exe/fetch.php?media=struktura_bazy.svg

[52] Można posiłkować się wykazem zmiennych przygotowanych dla pakietu ZPD dla R (http://zpd.ibe.edu.pl/doku.php?id=r_zmienne), jednak analogiczny wykaz dla wszystkich kolumn w fizycznej strukturze bazy danych nie istnieje.