

Metoda wstępnej analizy polegająca na tworzeniu słowników metadanych w projektach budowy analitycznych hurtowni danych

Rafał Renk

Akademia Techniczno-Rolnicza, Bydgoszcz

ITTI Sp. z o. o., Poznań

e-mail: renk@itti.com.pl

Andrzej Adamczyk

ITTI Sp. z o. o., Poznań

e-mail: adamczyk@itti.com.pl

prof. Witold Hołubowicz

Uniwersytet im. Adama Mickiewicza w Poznaniu

ITTI Sp. z o. o., Poznań

e-mail: holub@itti.com.pl

Abstrakt

Artykuł przedstawia proces tworzenia słowników metadanych jako metodę stosowaną w ramach wstępnego etapu projektu budowy analitycznych hurtowni danych. Utworzone słowniki metadanych stanowią ontologię analizowanych obszarów informacyjnych, a także są doskonałym punktem wyjścia do stworzenia interfejsu dostępu do danych w hurtowni. Efektywność takiego podejścia jest szczególnie duża w przypadku, kiedy dane zasilające hurtownię pochodzą z wielu obszarów informacyjnych oraz występuje wiele odmiennych interpretacji ich znaczenia. Korzyścią z zastosowania opisanego metody, oprócz ujednoczenia semantyki danych pochodzących z różnych obszarów informacyjnych, jest przedstawienie danych na jednym poziomie szczegółowości wraz z określeniem zależności i hierarchii. Opis metody poparty jest przykładem jej zastosowania w projekcie budowy ogólnopolskiej analitycznej hurtowni danych dla instytucji administracji publicznej. Przykład ten obejmuje metadane odzwierciedlające logikę danych, a nie ich parametry fizyczne.

1. Wprowadzenie

Ilość danych generowana przez obecnie eksploatowane w instytucjach i firmach systemy ERP (ang. *Enterprise Resource Planning*), CRM (ang. *Customer Relationship Management*), SCM (ang. *Supply Chain Management*) i inne systemy operacyjne jest ogromna. Tak duża ilość dostępnych danych oraz potrzeby analityczne kadry kierowniczej, która w wyniku analizy tych danych może otrzymać cenne informacje dotyczące sytuacji instytucji/firmy i perspektyw jej rozwoju spowodowały potrzebę budowy systemów informatycznych realizujących tego rodzaju zadania. Takimi systemami są analityczne hurtownie danych. Budowa takich systemów nie jest zadaniem łatwym i wymaga dobrego zrozumienia potrzeb biznesowych instytucji, dla której taki system miałby być tworzony. Poza potrzebami biznesowymi przy budowie tego rodzaju systemów istotne jest dobre rozpoznanie dostępnych danych w istniejących systemach informatycznych oraz określenie powiązań pomiędzy tymi danymi, co pozwala osiągnąć pełną interpretowalność danych podnosząc ich jakość a w efekcie gwarantuje wymierną korzyść z budowy hurtowni. Przedstawiona w artykule **metoda wstępnej analizy danych** wychodzi naprzeciw tym potrzebom.

W opracowaniu przedstawiono podstawowe pojęcia oraz relacje pomiędzy tymi pojęciami wykorzystywane w proponowanej metodzie wstępnej analizy danych. Do pojęć tych należą m.in.: dane, metadane, ontologia i analityczna hurtownia danych. Następnie przedstawiony został opis samej metody wstępnej analizy danych wykorzystanej przy realizacji projektu budowy analitycznej hurtowni danych. Na zakończenie artykułu omówione są korzyści, jakie wynikają z wykorzystania proponowanej metody na bazie doświadczeń zdobytych podczas projektu budowy ogólnopolskiej analitycznej hurtowni danych dla instytucji administracji publicznej.

2. Dane, metadane i ontologie

2.1. Dane i metadane

Dane w systemie informatycznym to informacje zapisane w pewnym obszarze pamięci komputera. Dane mogą stanowić pojedynczą informację np. imię lub nazwisko albo zespół powiązanych relacyjnie ze sobą informacji [WIEM]. Same dane jednak bez podania kontekstu ich wykorzystania niewiele znaczą dla użytkownika systemu informatycznego. Dopiero opisanie danych (określenie ich kontekstu) poprzez metadane nadaje znaczenie tym danym.

Jedną z najpopularniejszych definicji metadanych jest definicja określająca metadane jako „*dane o danych*” [KRMW98, Swet00]. Nie istnieje jednoznaczne rozróżnienie pomiędzy daną i metadaną. Metadane mogą stanowić dane dla kolejnych metadanych co prowadzi do budowy hierarchii metadanych (niekiedy tego rodzaju metadane określane są jako metametadane). Inną definicją metadanych jest definicja mówiąca, że metadane to: „*suma wszystkiego, co ktoś może powiedzieć o dowolnym obiekcie informacyjnym na dowolnym poziomie agregacji*” [Swet00]. W kontekście tej definicji obiekt informacyjny jest rozumiany jako cokolwiek, co może być zadresowane i manipulowane przez człowieka lub system jako byt dyskretny¹. W ogólności obiekt informacyjny, niezależnie od formy jaką może przyjąć, posiada trzy cechy:

- zawartość (ang. *content*) – odnosząca się do tego, co obiekt zawiera lub co przedstawia,
- kontekst (ang. *context*) – wskazanie aspekty związane z: kto, co, dlaczego, gdzie, jak, związane z obiektem,

¹ Należy tu podkreślić, że metadane nie koniecznie muszą mieć postać cyfrową; metadane mogą być przetwarzane do postaci cyfrowej.

- struktura (ang. *structure*) – odnosi się do formalnego zbioru powiązań wewnątrz lub pomiędzy obiektami informacyjnymi.

Metadane stanowią informację wykorzystywaną w celu m.in.: identyfikacji i reprezentacji zasobów, zapewnienia współpracy między różnymi systemami, wspomaganie zarządzania systemem informatycznym, określonego wykorzystania danych przechowywanych w systemie [Swet00]. Metadane wykorzystywane są w wielu dziedzinach m.in. w: zdalnym nauczaniu (w jego różnych aspektach m.in. opisie kursu, organizacji pytań itp.)², bibliotekach cyfrowych³, sieci Internet⁴, systemach informacji przestrzennej GIS⁵, bazach chorób w medycynie, hurtowniach danych, wyszukiwarkach internetowych itp.

W odniesieniu do hurtowni danych pojęcie to jest definiowane jako „wszystkie informacje w środowisku hurtowni danych, które nie są samymi danymi” [KRMW98]. W związku z dużą ilością różnych dodatkowych informacji w hurtowni danych nie będącymi samymi danymi dokonano dalszego podziału tych informacji na metadane „back room” i metadane „front room” [KRMW98]. Metadane *back room* związane są z procesami ETL (ang. *extract, transform, load*). Metadane *front end* są bardziej opisowe i korzystają z nich głównie użytkownicy końcowi i administratorzy hurtowni. Inny podział metadanych w odniesieniu do hurtowni danych wyróżnia:

- metadane techniczne - związane z procesami ETL,
- metadane operacyjne/administracyjne - związane z użytkowaniem i utrzymywaniem hurtowni danych,
- metadane biznesowe odnoszące się do opisu danych przeznaczonych dla użytkownika końcowego.

2.2. Ontologie

Wykorzystywane w systemach informatycznych metadane mogą przybierać dowolną formę opisu i korzystać z dowolnego zbioru pojęć. W celu utrzymania opisywanych informacji w sposób zorganizowany konieczne jest zapewnienie pojęciom wykorzystywanym przez metadane w systemach informatycznych spójności. Spójność taka pozwala na przetwarzanie metadanych przez maszyny i określana jest terminem słowników z kontrolowanymi pojęciami. Taka struktura słownictwa wykorzystywana przez metadane pozwala na ich organizację na wiele różnych sposobów odzwierciedlających informacje występujące w rzeczywistości. Do sposobów tych zaliczamy m.in.:

- taksonomie (ang. *taxonomies*) – zbiór kontrolowanych pojęć, typowo w układzie hierarchicznym np. Polska, woj. małopolskie, Zakopane,
- tezaury (ang. *thesaurus*) – są to taksonomia, które dodatkowo zawierają odniesienia do synonimów i zwrotów bliskoznacznych związanych z definiowanym pojęciem,
- ontologie (ang. *ontology*) – są podobne do tezaury ale wykorzystują bogatsze semantycznie relacje pomiędzy pojęciami i atrybutami oraz ściśle określone reguły specyfikacji pojęć i relacji. Ponieważ ontologie specyfikują więcej niż tylko pojęcia i mogą być automatycznie czytane przez maszyny wykorzystywane są do reprezentacji wiedzy (ang. *knowledge representation*).

Taksonomie i tezaury stanowią również swego rodzaju „uproszczone” ontologie, które wykorzystują tylko niektóre z możliwości oferowanych przez zasady konstrukcji ontologii.

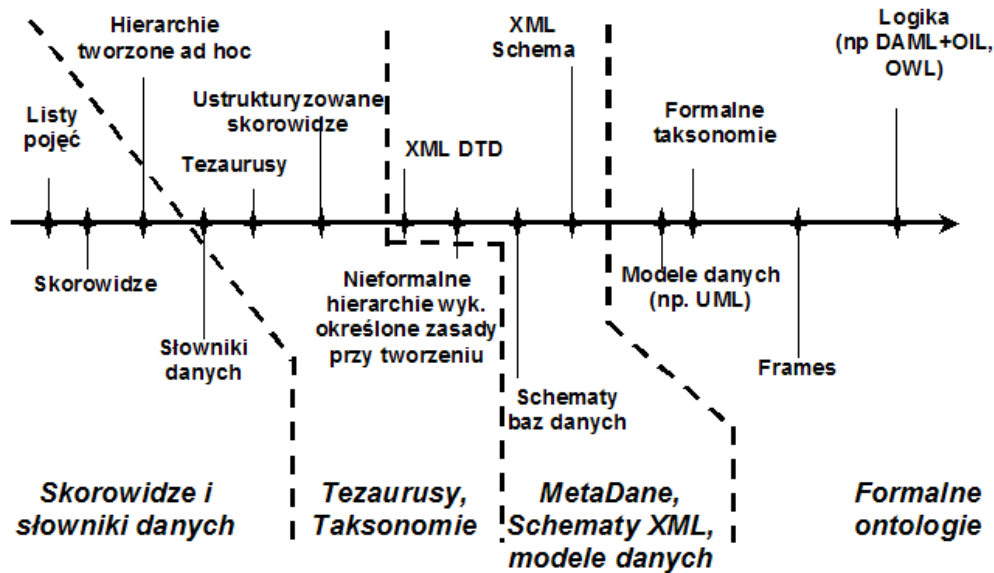
² np. specyfikacje konsorcjum IMS

³ np. specyfikacja Dublin Core

⁴ specyfikacje grupy roboczej Semantic Web Activity w ramach organizacji W3C (m.in.: OWL, RDF)

⁵ np. specyfikacje organizacji OpenGIS

Rys. 1 przedstawia rozwój ontologii w ostatnich latach. Na osi poziomej przedstawiony jest czas i ewolucja ontologii od skorowidzów i słowników danych poprzez taksonomie i tezaury, modele danych do formalnych ontologii.



Rys. 1. Rodzaje ontologii

Z historycznego punktu widzenia ontologia definiowana jest jako "nauka, która bada byt jako taki i przysługujące mu atrybuty istotne. Nie jest ona żadną z tzw. nauk szczegółowych, bo żadna z tych nauk nie bada ogólnie bytu jako takiego, lecz wyodrębnia pewną część bytu i bada jego własności"⁶.

Stosowana w informatyce definicja słowa ontologia to specyfikacja konceptualizacji. Konceptualizacja to z kolei obiekty, koncepty oraz powiązania, które definiują pewien obszar wiedzy z dziedziny problemu, który możemy nazwać światem. Ontologia jest jawną specyfikacją konceptualizacji.

Ontologia definiuje wykorzystywane pojęcia do opisu i reprezentacji określonego obszaru informacyjnego. Obszar informacyjny stanowi w tym przypadku określony obszar dziedziny lub wiedzy jak: medycyna, naprawa samochodów, ekonomia, itp. Z każdym obszarem informacyjnym związane są określone dane stanowiące obszar danych. Te same dane mogą należeć do różnych obszarów informacyjnych. Ontologie zawierają również definicje podstawowych pojęć w określonej dziedzinie i relacje między nimi wykorzystywane przez komputer. Pozwalają one na zapis wiedzy w określonej dziedzinie jak również wiedzy obejmującej wiele dziedzin. Ontologie mogą być wykorzystywane przez ludzi, bazy danych i aplikacje, które współdzielą pewien obszar informacyjny. W odróżnieniu od schematów baz danych ontologia:

- wykorzystuje język do definiowania pojęć i relacji składniowo i znaczeniowo bogatszy niż języki wykorzystywane przy projektowaniu schematów w bazach danych,
- opisuje wykorzystywaną informację w sposób podobny do tekstów zapisanych w języku naturalnym - nie w postaci tabelarycznej,
- bazuje na ogólnie przyjętej terminologii, posiada możliwości współdzielenia i wymiany informacji z innymi systemami opartymi na ontologii,
- dostarcza informacji na temat analizowanej domeny a nie tylko struktury danych.

⁶ Arystoteles, Metafizyka, III, 1, 1003a

2.3. Analityczne hurtownie danych

Generalnie systemy analityczne to takie systemy, które dostarczają informacje wykorzystywane do analizy problemu lub sytuacji. W ramach systemów analitycznych wykonywane jest przetwarzanie analityczne rozumiane jako przetwarzanie danych w celu wspomaganie decyzji strategicznych i związanych z zarządzaniem instytucją. Podstawą przetwarzania analitycznego są porównania oraz analizy wzorców i trendów na podstawie dużej ilości danych z szerokiego przedziału czasowego [PKB99]. Hurtownia danych pełni rolę analitycznej bazy danych, która stanowi podstawę systemu wspierającego proces podejmowania decyzji (ang. *Decision Support System - DSS*) i systemu informacyjnego dla kierownictwa (ang. *Executive Information System - EIS*). Oznacza to, że hurtownia taka powinna wspierać potrzeby informacyjne i analityczne w szerokim zakresie danych, poprzez dostarczanie zintegrowanych i odpowiednio przetworzonych danych historycznych obejmujących całą instytucję w sposób umożliwiający przeprowadzenie analizy. W celu przeprowadzenia samej analizy w oparciu o dane przechowywane w hurtowni danych powstało wiele narzędzi, które w sposób przyjazny dla użytkownika pozwalają na dostęp do przechowywanych.

Tradycyjne systemy relacyjne są projektowane dla aplikacji dokonujących przetwarzania transakcyjnego i realizacji prostych zapytań. W ogólności w systemach takich dane przechowywane w systemie to dane, jakie spodziewamy się otrzymać wykonując zapytanie. Systemy takie nie są projektowane z myślą o analizie wielowymiarowej uwzględniającej przebiegi czasowe.

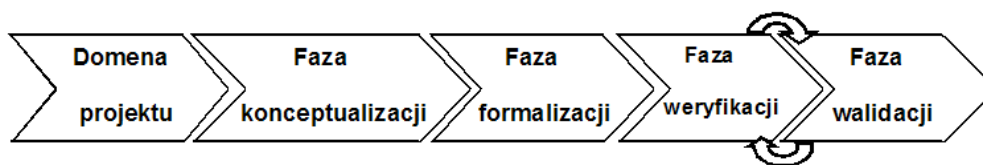
Analityczna hurtownia danych operuje na strukturach danych wielowymiarowych obejmujących szeroki zakres danych i wiele obszarów informacyjnych dotyczących instytucji. Hurtownia taka charakteryzuje się możliwością: importu danych źródłowych, odczytu danych z hurtowni, zapisu wyników dokonanych analiz i symulacji jak również brakiem wysokich wymagań odnośnie dokładności, kompletności i aktualności przechowywanych danych. W celu przeprowadzenia określonych analiz i symulacji nie jest wymagane posiadanie najaktualniejszych danych ze 100% dokładnością jak ma to miejsce np. w systemach księgowych. Hurtownia danych zaprojektowana z myślą o przeprowadzaniu analiz musi uwzględniać możliwość współpracy z różnymi narzędziami do przeprowadzania analiz, koniecznością generowania niestandardowych raportów przygotowywanych w całości przez użytkowników systemu oraz konieczności integracji metadanych pomiędzy hurtownią danych i wykorzystywanymi narzędziami analitycznymi w celu zapewnienia jednolitego rozumienia danych zawartych w hurtowni.

Analityczna hurtownia danych to system informatyczny, w skład, którego wchodzi: hurtownia danych wykorzystywana głównie jako źródło odpowiednio przetworzonych danych dla narzędzi wspomagających wykonywanie przetwarzania analitycznego. Całość systemu informatycznego powinna uwzględniać funkcjonalność pozwalającą na korzystanie i administrowanie jednym zbiorem metadanych dla wszystkich wykorzystywanych elementów systemu.

3. Opis metody wstępnej analizy danych

Kiedy klient podejmuje decyzję o wdrożeniu hurtowni danych, okazuje się, że posiada on już wiele systemów, które można nazwać tematycznymi hurtowniami danych lub centralnymi bazami danych systemów operacyjnych, gdzie zgromadzone są użyteczne dane. Zdarza się, że klient nalega, aby użyć tych danych, jako źródła dla nowej hurtowni (mimo, że jest to sprzeczne z obecnie obowiązującą architekturą hurtowni danych). W takich przypadkach etap projektowy wdrożenia hurtowni warto poprzedzić etapem wstępnej analizy semantycznej istniejących i potrzebnych danych. Analiza ta w odróżnieniu od analizy techniczno-projektowej powinna abstrahować od struktury danych w docelowej bazie i od względów implementacyjnych. Przeprowadzenie takiego etapu jest szczególnie użyteczne w przypadku dużych firm i instytucji, gdzie rozumienie znaczenia poszczególnych danych może się znacznie różnić na przykład pomiędzy różnymi działami i departamentami.

Rys. 2 przedstawia fazy realizacji metody wstępnej analizy danych dla wybranego obszaru informacyjnego.



Rys. 2. Fazy realizacji wstępnej analizy danych

W ramach poszczególnych faz wykonywane jest:

- domena projektu – analiza natury obszaru informacyjnego, określenie i skupienie się na głównych pojęciach związanych z analizowanym obszarem informacyjnym, analiza doświadczeń z poprzednio wykonywanych prac w istotnych obszarach informacyjnych (doświadczenia: zamawiającego, krajowe i światowe),
- faza konceptualizacji – wstępne opracowywanie struktury tworzonych słowników metadanych (ontologii), specyfikacja poziomu abstrakcji opisu danych, rozpoznanie i analiza danych, zebranie danych, utworzenie taksonomii dla opisywanych danych.
- faza formalizowania – określenie systematyki danych i organizacji powiązań, formalizowanie opisu koncepcji, dodawanie klas i relacji pomiędzy nimi, opracowanie specyfikacji danych w odniesieniu do analizowanego obszaru informacyjnego,
- faza weryfikacji – sprawdzenie i weryfikacja poprawności utworzonych słowników metadanych,
- faza walidacji – walidacja opracowanej specyfikacji danych (modelu danych), rozszerzenie i korygowanie utworzonych słowników metadanych.

W poszczególnych fazach realizacji metody wstępnej analizy danych w celu pozyskiwania informacji na temat analizowanego obszaru informacyjnego można wykorzystać m.in.:

- spotkania i wywiady bezpośrednie:
 - z ekspertami merytorycznymi analizowanych obszarów informacyjnych,
 - z informatykami odpowiedzialnymi za istniejące systemy informatyczne zawierające dane związane z analizowanymi obszarami informacyjnymi,
- uczestnictwo w spotkaniach i seminariach z przyszłymi użytkownikami systemu,
- przegląd istniejącej i historycznej dokumentacji związanej z zakresami analizowanych danych,
- przegląd doświadczeń krajowych i międzynarodowych we wdrażaniu systemów informatycznych związanych z analizowanymi obszarami danych,
- analiza aktów prawnych związanych z analizowanymi obszarami informacyjnymi,
- analiza publicznie dostępnych materiałów na temat znaczenia analizowanych danych,
- wykorzystanie wiedzy eksperckiej konsultantów.

Metoda wstępnej analizy danych w przeciwieństwie do metodyki etapowej charakteryzującej się równoległym prowadzeniem prac w poszczególnych obszarach informacyjnych, przyjmuje postępowanie inkrementacyjne. Ma to szczególne znaczenie gdy dane rozłożone są w różnych systemach źródłowych oraz przy krótkich terminach pracy ponieważ daje możliwość pełnej implementacji określonego zakresu danych. Postępowanie inkrementacyjne charakteryzuje się tym, że wszystkie opisane wyżej fazy (Rys. 2) wykonywane są dla wybranego pierwszego obszaru in-

formacyjnego, a następnie powielane w innych obszarach informacyjnych. Rozszerzanie zakresu informacyjnego systemu powoduje niekiedy konieczność dokonywania zmian w opracowanych już w trakcie analizy obszarów informacyjnych (faza walidacji i weryfikacji na Rys. 2). Dla każdego obszaru informacyjnego wykonana dokumentacja powinna zawierać:

- słownik pojęć dziedzinowych, związanych z analizowanym obszarem informacyjnym,
- opis słowny obszarów informacyjnych,
- model danych w postaci np. diagramów UML (ang. *Unified Modelling Language*) [OMG],
- słowniki metadanych – zawartość danych związanych z elementami modelu, ich hierarchie, itp. wraz ze wskazaniem na źródła danych dla poszczególnych elementów modelu metadanych,
- ewentualne uwagi dodatkowe.

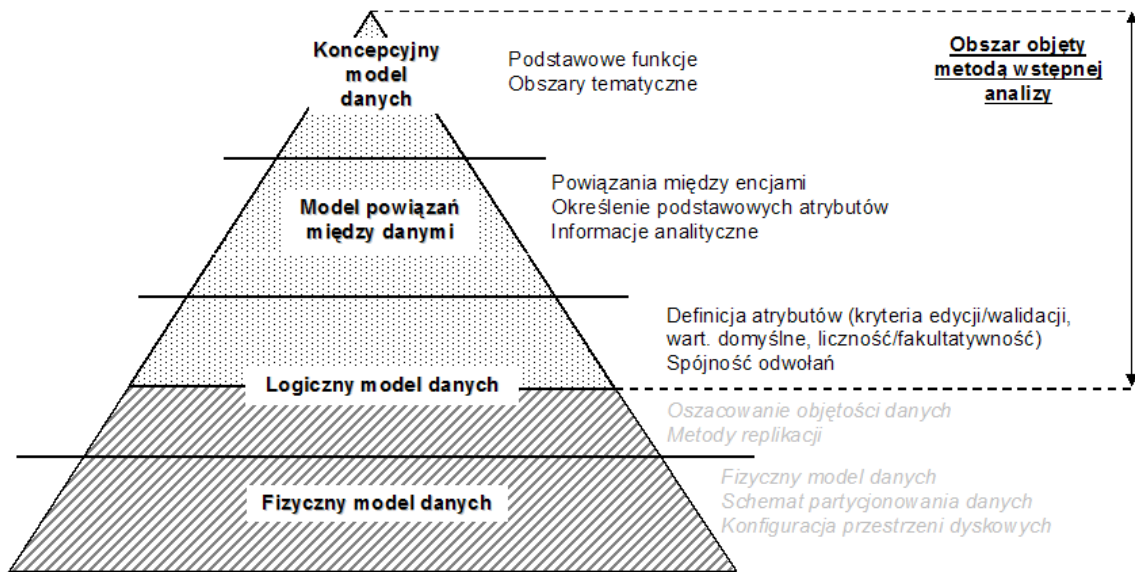
W oparciu o informacje zawarte w dokumentach składowych opisujących poszczególne obszary informacyjne opracowuje się dokument zbiorczy np. pt. „Ogólna Specyfikacja Danych”. Konsekwencją wybranej metodyki inkrementacyjnej jest fakt, że każdy z dokumentów ewoluuje w czasie całego projektu dostosowując się do wymogów stawianych przez kolejne porcje analizowanych danych (faza walidacji i weryfikacji na Rys. 2).

Jako pierwszy obszar informacyjny do analizy należy wybrać obszar o odpowiednio dużej złożoności. Dzięki temu początkowy model danych będzie wystarczająco złożony, aby analiza kolejnych obszarów danych wprowadzała jak najmniej modyfikacji do założonej struktury. Dodatkowo obszar ten powinien być stosunkowo dobrze opisany, co pomoże w przygotowaniu rzetelnych podwalin pod kolejne fazy inkrementacji analizy i opisu danych.

Dane powinny być opisane w sposób logiczny, a nie techniczny. Oznacza to, że wytworzone modele danych i słowniki powinny abstrahować zarówno od technologii, jak i metody implementacji systemu informatycznego, który miałby opierać się na tychże danych. Słowniki danych mają na celu jedynie przekazać **wiedzę** niezbędną do stworzenia modelu technicznego w procesie projektowania docelowego systemu informatycznego.

Dodatkowo większy nacisk należy położyć na wymagania użytkowe opisywanych danych niż na ich fizyczną lokalizację. W pierwszej fazie następować powinna identyfikacja potrzebnych danych, a dopiero w drugim kroku poszukuje się źródła tych danych. Podejście to pozwala na nie ograniczanie się do danych dostępnych w konkretnym momencie w systemach informatycznych. W efekcie jednak nie ma gwarancji, że dla każdego elementu danych zostanie wskazane konkretne źródło w bazie danych.

Metoda wstępnej analizy nie obejmuje zagadnień związanych z modelem fizycznym danych ani też dokładnego określenia wymagań przyszłych użytkowników systemu. Jej głównym celem jest przygotowanie zakresu danych pod względem pojęciowym do właściwego etapu analizy i projektu systemu. Zakres prac związanych z metodą wstępnej analizy danych przedstawiony jest na Rys. 3.



Rys. 3. Zakres danych objęty metodą wstępnej analizy danych

4. Korzyści wynikające z zastosowania metody analizy wstępnej

Korzyści płynące z zastosowania metody wstępnej analizy danych przedstawione są na podstawie doświadczeń wynikających z zastosowania tej metody w projekcie budowy ogólnopolskiej analitycznej hurtowni danych dla instytucji administracji publicznej. W projekcie tym wykorzystano metodę wstępnej analizy danych do opracowania słowników metadanych stanowiących ontologię analizowanych obszarów informacyjnych. W celu formalnego zapisu ontologii posłużono się diagramami UML (ang. *Universal Modeling Language*) [OMG].

Do głównych korzyści wynikających z zastosowania metody wstępnej analizy danych zaliczamy:

- podnoszenie jakości danych w wyniku dokładnego określenia znaczenia danych i powiązań z innymi danymi w projektowanej hurtowni danych - ujednolicenie znaczenia wykorzystywanych pojęć pochodzących z różnych systemów informatycznych (systemów źródłowych) oraz powiązań pomiędzy danymi reprezentowanymi przez te pojęcia,
- przedstawienie danych na jednym poziomie abstrakcji zrozumiałym dla osób odpowiedzialnych za aspekty merytoryczne w analizowanych obszarach informacyjnych,
- możliwość realizacji metody analizy wstępnej równoległe z opracowywaniem zakresu i planu projektu pozwalające:
 - zamawiającemu na lepsze zrozumienie własnych potrzeb względem projektowanego systemu informatycznego,
 - przyszłemu wykonawcy na przyspieszenie realizacji etapu analizy,
- określenie źródeł danych dla obszarów informacyjnych oraz wstępna weryfikacja zakresu danych dostępnych w źródłach danych (dostępność danych w systemach źródłowych),
- wykorzystanie opracowanych słowników metadanych jako punkt wyjścia do stworzenia interfejsu dostępu do danych w hurtowni.

Z punktu widzenia analitycznej hurtowni danych istnieje konieczność zapewnienia dobrej jakości danych. Obecnie według opinii ekspertów zła jakość danych była przyczyną porażki ponad 70% projektów hurtowni danych⁷. Na jakość danych składają się takie cechy jak:

- dostępność – czy dane są dostępne dla każdego z użytkowników, który ich potrzebuje i przez odpowiedni czas?
- interpretowalność – czy istnieją metadane opisujące składnię i znaczenie danych?
- użyteczność – czy dane są adekwatne, aktualne, łatwe do manipulowania nimi?
- wiarygodność – czy dane są zgodne z prawdą, kompletne, spójne (zgodne) wewnątrznie, na wymaganym poziomie dokładności, pochodzą z wiarygodnego źródła?

Najczęstsze problemy związane z jakością danych w procesie tworzenia hurtowni to:

- brak metadanych (na pierwszym miejscu!),
- brak samych danych,
- nieuporządkowana struktura danych (dane są tam, gdzie nie powinny być),
- brak historyczności w momencie tworzenia hurtowni (później oczywiście powstanie ona automatycznie w hurtowni),
- brak aktualności danych,
- niski stopień wiarygodności danych,
- brak możliwości przyrostowego dodawania danych do hurtowni,
- niespójność danych (występowanie danych nie zawartych w słownikach; dane ze słowników, ale stanowiące połączenie nie występujące w rzeczywistości),
- niejednoznaczność np. dwie wartości opisując tę samą zmienną.

W kontekście powyższego opisu praca nad słownikami metadanych (ontologiami obszarów informacyjnych) w ramach wstępnej analizy danych jest bardzo ważna dla zapewnienia interpretowalności, użyteczności i wiarygodności danych, a przez to przyspieszenia i zmniejszenia ryzyka projektu budowy hurtowni danych. Nie przeprowadzenie wstępnej analizy danych może prowadzić do:

- zwiększenia kosztów budowy hurtowni,
- pominięcia pewnego zakresu danych przy budowie hurtowni.

Z punktu widzenia metody wstępnej analizy ważne jest aby zapewnić mechanizm utrzymywania słowników metadanych oraz przeprowadzanie wzajemnej weryfikacji i uzupełniania słowników korzystając z wielu źródeł tych samych danych.

Metoda wstępnej analizy pozwala na przedstawienie danych na jednym poziomie szczegółowości (wysoki poziom abstrakcji) oraz ujednoliconym znaczeniu poszczególnych pojęć pochodzących z różnych obszarów informacyjnych. Takie podejście do opisu danych pozwala na przedstawianie informacji pochodzących z różnych źródeł w sposób spójny i logiczny. Nie zawsze jest to możliwe przy wykorzystaniu istniejących słowników związanych z tematycznymi systemami źródłowymi gdzie jedno pojęcie może być rozumiane w każdym systemie źródłowym w różny sposób. Dodatkowo przedstawienie danych na wysokim poziomie abstrakcji pozwala na łatwiejsze ich zrozumienie oraz weryfikację przez osoby odpowiedzialne za aspekty merytoryczne w analizowa-

⁷ metoda wstępnej analizy danych nie obejmuje bezpośrednio zapewnienia jakości danych w procesach ETL, tylko na poziomie interpretowalności, użyteczności i wiarygodności danych.

nych obszarach informacyjnych. Przedstawienie i krótkie omówienie diagramu klas zapisanego w języku UML pozwalało osobom merytorycznym na szybką weryfikację przedstawianego zapisu.

Metoda wstępnej analizy może być realizowana równolegle z opracowywaniem zakresu i planu projektu dla przyszłego wykonawcy systemu. Pozwala to zamawiającemu na lepsze zrozumienie własnych potrzeb biznesowych względem projektowanego systemu informatycznego, w tym identyfikacja obszarów informacyjnych (definicja modelu informacyjnego) oraz wstępne określenie zakresu i atrybutów odnoszących się do zidentyfikowanych obszarów informacyjnych (logiczny model danych). Wstępna analiza danych pozwala również wykonawcy na szybsze przejście do etapu projektowania struktur danych i mechanizmów aplikacyjnych bez konieczności zdobywania złożonej i rozproszonej wiedzy o danych potrzebnych do zasilenia hurtowni. Przygotowane ontologie stanowią też dobry punkt wyjścia do zapoznania się analityka ze strony wykonawcy z analizowanym obszarem informacyjnym przed spotkaniem z użytkownikiem końcowym. Dodatkowo powiązania pomiędzy poszczególnymi pojęciami w utworzonej ontologii stanowią dobry punkt do rozpoczęcia projektowania i budowy modułów ETL.

Wynik metody wstępnej analizy danych pozwala na logiczne (np. dokumenty, formularze, raporty) jak i fizyczne (istniejące systemy informatyczne) określenie źródeł danych dla analizowanych obszarów informacyjnych. Metoda ta pozwala na wstępną weryfikację zakresu danych dostępnych w poszczególnych systemach źródłowych.

Przygotowane w ramach wstępnej analizy danych ontologie stanowią dobry punkt do rozpoczęcia projektowania i budowy interfejsu użytkownika uwzględniającego możliwość przedstawienia znaczenia danych przechowywanych w hurtowni. Tego typu interfejs jest swego rodzaju definicją perspektywy, z której użytkownik patrzy na dane poprzez ich logiczne znaczenie w różnym kontekście np. prawnym, praktycznym, ekonomicznym.

5. Podsumowanie

Artykuł ten przedstawia opis metody wstępnej analizy danych polegająca na budowie słowników metadanych oraz korzyści wynikające z praktycznego zastosowania tej metody w projekcie analitycznej hurtowni dla instytucji administracji publicznej. Doświadczenia te pokazują dużą praktyczną przydatność zastosowania tej metody. Utworzone słowniki metadanych m.in.:

- zapewniają źródło informacji o danych dostępnych w hurtowni dla narzędzi analitycznych i użytkowników korzystających z tych danych,
- są uwiarygodnieniem analizowanych danych i potwierdzenie ich aktualności,
- umożliwiają potwierdzenie dobrej jakości analizowanych danych.

Metoda ta ma szczególne znaczenie w przypadku dużych firm i instytucji, gdzie zrozumienie znaczenia poszczególnych danych może się znacznie różnić na przykład pomiędzy różnymi działami oraz gdy klient nalega aby użyć jako źródeł danych istniejących systemów informatycznych. Nie oznacza to, że dla innego rodzaju projektów hurtowni danych metoda ta może znaleźć zastosowanie.

Ponadto rozpoczęcie budowy repozytorium metadanych jest świadectwem doceniania ważności zadania porządkowania informacji wewnątrz instytucji, polegającej na: inwentaryzacji danych istniejących w systemach transakcyjnych, ustaleniu jednego źródła (jednej „prawdy”) wiarygodnej informacji oraz ujednoczeniu znaczenia poszczególnych pojęć wykorzystywanych przez różne jednostki administracyjne, zasad weryfikacji jakości i transformacji danych na etapie zasilania hurtowni danych, ustalenie harmonogramu migracji danych do hurtowni itd.

Dobrze zaprojektowana hurtownia danych to podstawa dla systemów analitycznych i źródło informacji w podejmowaniu decyzji przez kadrę kierowniczą w przedsiębiorstwa. Błędy w projekcie

hurtowni i nie zapewnieniu interpretowalności, użyteczności i wiarygodności danych mogą spowodować załamanie całego systemu informowania w instytucji. Późniejsze usuwanie powstałych błędów może być natomiast bardzo kosztowne.

Bibliografia

- [KRMW98] Kimball R., Reeves L., Margy R., Warren T.: The Data Warehouse Lifecycle Toolkit, Wiley & Sons, 1998, ISBN 0-471-25547-5
- [Swet00] Gilliland-Swetland A. J.: Introduction to metadata, www.getty.edu/research/institute/standards/intrometadata, 2000
- [PKB99] Poe V., Klauer P., Brobst S.: Tworzenie hurtowni danych, WNT, 2000, ISBN 83-204-2435-6
- [OMG] Specyfikacja języka UML opracowana przez organizację OMG, <http://www.uml.org/>
- [WIEM] Internetowy portal wiedzy WIEM, <http://portalwiedzy.onet.pl/>