Leopoldo Bertossi1⋆⋆, Flavio Rizzolo2⋆⋆⋆, Lei Jiang3

1 Carleton University, Ottawa, Canada. bertossi@scs.carleton.ca

2 Carleton University, Ottawa, Canada. flavio@scs.carleton.ca

3 University of Toronto, Toronto, Canada. [leijiang@cs.toronto.edu](mailto:leijiang@cs.toronto.edu)

**Streszczenie** Uzasadniamy, formalizujemy i badamy takie pojęcia jak ocena jakości danych oraz odpowiadanie na zapytania dotyczące jakości danych jako działania zależne od kontekstu. Konteksty dla oceny i wykorzystania podręcznych źródeł danych są modelowane jako zewnętrzne bazy danych (które mogą być zmaterializowane albo wirtualne) i mapowania wewnątrz zbiorów oraz podręczne źródła danych. W ten sposób, kontekst staje się "uzupełnieniem" źródła danych odnośnie do systemu integracji danych. Zaproponowany model pozwalana na naturalne rozszerzenia, jak branie pod uwagę orzeczeń o jakości danych, a nawet bardziej wyraziste ontologie dla oceny jakości danych.

**Tematy** Jakość i oczyszczanie danych.

Jakość danych jest zależna od kontekStu

# Wstęp

Oszacowanie jakości źródeł danych jest zależne od kontekstu, tj. pojęcia „dobrych” lub „złych” danych nie mogą być oddzielone od kontekstu, w jakim dane są wytwarzane lub stosowane. Na przykład, dane o rocznej sprzedaży danego produktu z wahaniami sezonowymi mogą być uznane za dane dotyczące jakości przez analityka biznesowego oszacowującego roczny przychód produktu. Jednakże, te same dane mogą nie być wystarczająco dobre dla kierownika magazynu, który stara się oszacować zamówienia na następny miesiąc.

Ponadto, jakość danych jest związana z rozbieżnością pomiędzy rzeczywistymi przechowywanymi wartościami i "prawdziwymi" wartościami, które miały być lub oczekiwane jest że mają być przechowywane. Na przykład, jeżeli pomiar temperatury jest dokonywany z użyciem wadliwego termometru, zapisana wartość (pomiar) różniłaby się od właściwej wartości (rzeczywista temperatura), która była rzekomo tą jedną która miała być przechowywana. Jest to przykład *semantycznie nieprawidłowych danych* [3].

Ponadto, inny rodzaj semantycznej rozbieżności występuje wtedy, gdy *zmysły lub  
znaczenie* przypisywane przez różne środki do rzeczywistych wartości w bazie danych  
nie zgadzając się [19], jak pokazano w przykładzie 1. W tym artykule skupimy się na jakości danych (DQ – Data Quality) problemów powodowanych przez ten typ semantycznej rozbieżności.

*Przykład 1.* Tom jest pacjentem w szpitalu. Kilka razy na dzień jego temperatura jest mierzona i zapisywana przez pielęgniarkę. Jego lekarz, John chce zobaczyć temperaturę Toma codziennie około południa, aby śledzić jej zmiany. Informacja, której John potrzebuje pojawia się w relacji TempNoon w Tabeli 1, która zawiera temperatury pomiędzy 11:30 a 12:30 w ciągu dnia dla każdego z pacjentów Johna.

**TempNoon**

|  |  |  |  |
| --- | --- | --- | --- |
| **Pacjent** | **Wart.** | **Czas** | **Data** |
| Tom Waits | 38.5 | 11:45 | Sep/5 |
| Tom Waits | 38.2 | 12:10 | Sep/5 |
| Tom Waits | 38.1 | 11:50 | Sep/6 |
| Tom Waits | 38.0 | 12:15 | Sep/6 |
| Tom Waits | 37.9 | 12:15 | Sep/7 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Pacjent** | **Wart.** | **Czas** | **Data** |
| Tom Waits | 38.5 | 11:45 | Sep/5 |
| Tom Waits | 38.0 | 12:15 | Sep/6 |
| Tom Waits | 37.9 | 12:15 | Sep/7 |

John ma dodatkowe wymagania co do *jakości* pomiaru temperatury jego pacjentów: muszą one być przeprowadzane przez certyfikowane pielęgniarki z termometrem doustnym.5 września, nieświadoma nowych wymagań, Cathy mierzy temperaturę Toma o 12:10 za pomocą termometru *na podczerwień* i zapisuje wynik jako liczbę (krotkę) 2 w tabeli 1. Ponieważ wykorzystywany przyrząd nie pojawia się w widoku, który John ma w danych, interpretuje on wartość 38.2oC jako zmierzoną za pomocą termometru doustnego.

Jest to przykład rozbieżności pomiędzy semantyką wartości w sposób przewidziany przez producenta danych (38.2oC zmierzone za pomocą termometru na podczerwień) i semantyki oczekiwanej przez konsumenta danych (38.2oC zrobione za pomocą termometru doustnego). Ta krotka nie powinna pojawiać się w tabeli jakości, to znaczy w takiej, który spełnia wymagania Johna dotyczące jakości, ponieważ taka tabela będzie zawierać tylko temperatury zmierzone za pomocą termometru doustnego.

Podobny problem pojawia się w trzeciej krotce w Tabeli 1: Została zmierzona przez nową pielęgniarkę, Helen, która jeszcze nie jest certyfikowana, a tym samym nie spełnia jednego z wymagań lekarza. Ta krotka nie powinna pojawiać się w tabeli jakości zawierającej wyłącznie temperatury zmierzone przez dyplomowane pielęgniarki.

Tabela 2 rozwiązuje problemy związane z Tabelą 1 w odniesieniu do specyfikacji lekarza: Problematyczne druga i trzecia krotka nie pojawiają się w niej.

Jak możemy powiedzieć lub wierzyć że Tabela 2 zawiera wyłącznie dane jakościowe? Na pierwszy rzut oka nie różni się za wiele od Tabeli 1. Ta pozytywna ocena byłaby możliwa, jeśli mieliśmy bazę kontekstową zawierającą dodatkowe informacje, np. Tabele 3, 4 i 5.

|  |  |  |
| --- | --- | --- |
| **Data** | **Zmiana** | **Pielęg.** |
| Sep/5 | morning | Susan |
| Sep/5 | afternoon | Cathy |
| Sep/5 | night | Joan |
| Sep/6 | morning | Helen |
| Sep/6 | afternoon | Cathy |
| Sep/6 | night | Cathy |
| Sep/7 | morning | Susan |
| Sep/7 | afternoon | Susan |
| Sep/7 | night | Joan |

|  |  |  |
| --- | --- | --- |
| **Pielęg.** | **Data** | **Typ** |
| Susan | Sep/5 | Oral |
| Cathy | Sep/5 | Tymp |
| Joan | Sep/5 | Tymp |
| Helen | Sep/6 | Oral |
| Cathy | Sep/6 | Oral |
| Susan | Sep/7 | Oral |
| Joan | Sep/7 | Oral |

|  |  |
| --- | --- |
| **Imię** | **Rok** |
| Ann | 2003 |
| Cathy | 2009 |
| Irene | 2000 |
| Nancy | 1995 |
| Susan | 1996 |

Pierwsza relacja zawiera imię pielęgniarek na oddziale Toma Waitsa i zmiany na których pracują w ciągu dnia. Są to pielęgniarki robiące pomiary; ponieważ jest to mały oddział jest tylko jedna pielęgniarka na zmianę z tym zadaniem. Druga relacja rejestruje imię certyfikowanych pielęgniarek na oddziale, oraz rok w którym dostała certyfikat. Ostatnia relacja zawiera typ termometru którego każda pielęgniarka używa w ciągu dnia (na przykład doustny lub na podczerwień); każda pielęgniarka wykonuje wszystkie pomiary temperatury w ciągu dnia za pomocą tego samego rodzaju termometru. Ta informacja kontekstowa pozwala nam ocenić jakość danych w Tabelach 1 i 2.

W tym artykule bierzemy na poważnie intuicję i doświadczenie, że jakość danych jest zależna od kontekstu. Nasze sformalizowanie kontekstu podawane jest jako system zintegrowanych danych i metadanych, którego źródłem danych na podstawie oceny jakości jest szczególny i specjalny składnik. Dokładniej, kontekst dla oceny jakości danych w określonej instancji D schematu S jest podany przez wystąpienie I możliwie różnych schematów C, które mogłyby stanowić rozszerzenie S. W celu oceny jakości D, należy to "umieścić w kontekście", który został osiągnięty przez mapowanie D (oraz S) do kontekstowego schematu i danych. Właściwie, C może być bardziej skomplikowane, że jeden schemat lub instancja, a mianowicie zbiór schematów baz danych i instancji powiązanych ze sobą przez mapowania danych i schematu.

W naszym kontekście jakość instancji bazy danych D może być postrzegana jako "ślad" z kontekstowej, rozszerzonej bazy I. Ewentualnie dodatkowe informacje w I są tym co daje kontekst oraz wyjaśnia dane w schemacie D. Kontekstowy schemat i dane nie służą w danym przypadku do narzucenia jakości danych. Zamiast tego stosuje się: (A) Ocenianie jakości danych w instancji pod ręką; (B) Charakteryzowanie jakości odpowiedzi na pytania; oraz (c) Możliwość uzyskania tych odpowiedzi jakościowych do zapytania użytkownika. Wszystko to jest osiągnięte przez porównanie danej instancji D z instancją I, wirtualnej lub istniejącej, która może być zdefiniowana dla schematu kontekstowego na podstawie D źródeł zewnętrznych, które współpracują z danymi w schemacie kontekstowym i ewentualnych dodatkowych danych na poziomie kontekstowym, jak pokazano w przykładzie 1.

Powyższa instancja I może zostać zastąpiona przez znacznie bogatszy opis kontekstowy, na przykład pełnoprawnej ontologii. W tym stylu, ale ciągle w klasycznym scenariuszu bazy danych, możemy zdefiniować kilka dodatkowych *predykatów jakościowych* na C [19]. Mogą one być wykorzystane do oceny jakości danych w D (a także do jakości odpowiedzi zapytań z D, będziemy badać to później).

Następujące wkłady można znaleźć w tym artykule: (a) model kontekście oceny jakości danych. (b) jego stosowanie do czyszczenia lub odpowiadania na zapytania dotyczące jakości danych. (c) Jego zastosowanie do oceny jakości danych poprzez kilka naturalnych pomiarów, które powstają się z modelu. (d) Niektóre algorytmy dla wyżej wymienionych zadań w kilku konkretnych, ale wspólnych i naturalnych przypadkach. (e) stworzenie ram, które mogą być naturalnie rozszerzane w późniejszej pracy żeby objąć więcej ogólnej ontologii kontekstowych i zewnętrznie zdefiniowanych predykatów jakościowych.

Pozostała część artykułu jest zorganizowana w następujący sposób. W rozdziale 2 przedstawiono ogólne ramy kontekstowej jakości danych i zilustrowanie ich za pomocą działającego przykładu. W rozdziale 3, rozważamy dwa szczególne przypadki ogólnych ram, w których zakładamy, że mamy kontekstową instancję I, którą możemy wykorzystać do oceny jakości i przedstawić algorytm odpowiadający na zapytania dotyczące jakości w ramach tego założenia. W rozdziale 4, badamy bardziej skomplikowane przypadki, na przykład, gdzie nie istnieją takie przypadki kontekstowe. Omówimy podobne prace w pkt 5; i podsumujemy i odniesiemy się do na naszej bieżącej i przyszłej pracy w punkcie 6.

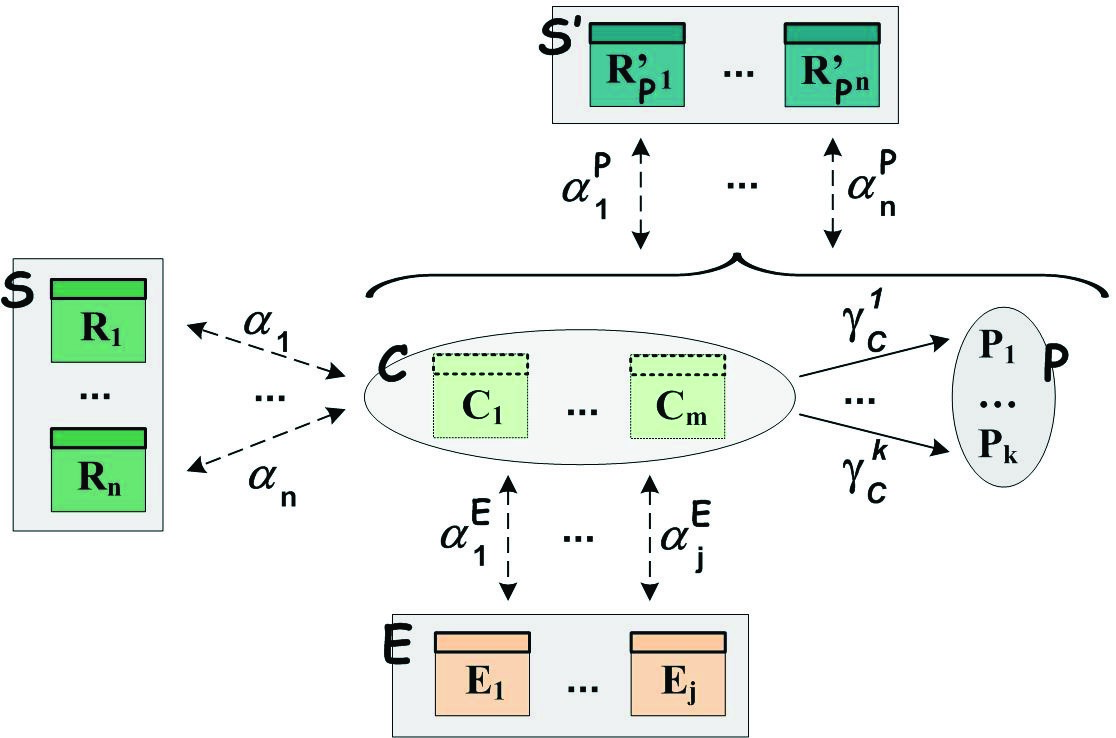
# Ramy dla jakości danych kontekstowych

Rozważmy relacyjny schemat S, z relacyjnymi predykatami *R,...* ∈ S. Ten Schemat określa język *L*(S) pierwszego rzędu logiki predykatów. W tym dokumencie bierzemy pod uwagę tylko zapytania jednostajne i widoki, na przykład koniunkcyjnych zapytań i ich związków, które zwykle piszemy w nie-rekursywnym Datalogu z wbudowaniami [1]. Rozważymy również instancję *D* z S, z rozszerzeniem *R*(*D*) dla R. itp. Jeśli instancje bazy danych są pomyślane jako skończone zbiory cząstek podstawowych, a następnie dla każdego *R* ∈ S, *R*(*D*) ⊆ *D*. Przypadki *D,R*(*D*)*,...* są to te podlegające oceny jakości w ramach systemu kontekstowego.

W najprostszej postaci, system składa się z kontekstowego relacyjnego schematu C, który może zawierać zbiór B wbudowanych predykatów. Możemy mieć lub nie instancję dla C. W bardziej skomplikowanym scenariuszu system kontekstowy może składać się z kilku schematów kontekstowych C1*,...,*C*n* a także zestawu zewnętrznych schematów E, które mogą być wykorzystywane przez system kontekstowy dla ocena instancji *D* od S.

W tych ogólnych ramach, uczestniczące schematy są powiązane przez *schemat mapowań*, podobnie jak te znajdujące się na przykład w systemach wirtualnych integracji danych (VDISs) [22, 4] lub wymiany danych [21], a nawet bardziej skomplikowane relacje logiczne, takie jak te, powszechne w systemach zarządzania danymi rówieśniczych [5, 6]. (Por [13] na podstawie analizy połączeń między tymi trzema obszarami.) Mapowania schematu przybierają formę korespondencji pomiędzy dwoma wzorami, jak zapytania lub widoki definicji, każdy z nich zawierające predykaty z jednego lub kilku schematów. W szczególności źródłem danych na podstawie oceny D będą mapowania do schematu kontekstowego. Powszechną formą połączenia lub mapowania jest z ∀*x*¯(*S*(¯*x*) → *φG* (¯*x*)), gdzie S jest relacyjnym predykatem źródła danych i *φ* (¯*x*) jest koniunkcyjnym zapytaniem na globalnym relacyjnym schemacie G. Te połączenie można znaleźć w *widokach lokalnych* (LAV) VDISs z otwartymi (lub dźwiękowymi) źródłami. Inną popularną formą połączenia jest forma ∀*x*¯(*ψ* (¯*x*) → *G*(¯*x*)), znajdujące się w global-as-view (GAV) VDISs ze źródłami otwartymi, gdzie *ψ* (¯*x*) jest koniunkcyjne zapytanie nad unią R z relacyjnych schematów u źródeł, i G jest globalnym relacyjnym predykatem. W *global-and-local-as-view* (GLAV) VDISs z otwartymi źródłami, znajdujemy skojarzenia między widokami (lub zapytaniami), z postaci ∀*x*¯(*ψ* (¯*x*) → *φ* (¯*x*)).

Rysunek 1 przedstawia ten ogólny scenariusz. Stosunki *Ri* w D podlegają ocenie jakości, która jest wykonywana za pomocą kontekstowego schematu C, która ma relacyjne predykaty *C*1*,...,Cm*. Istnieje również zbiór P *kontekstowych jakościowych predykatów* (CQPs) *P*1*,...,Pk*,, które są definiowane jako widoki C plus możliwie zewnętrznych źródeł *E*1*,...,Ej*.. W niektórych przypadkach, kombinacja schematów C, P, E może być postrzegana jako pojedynczy rozszerzony schemat kontekstowy. W innych przypadkach, to może być użyteczne, aby odróżnić je od siebie. Zewnętrzne predykaty *Ei* mogłyby mieć "pseudonimy" *Ei*′ w C, za pomocą prostych definicji jako mapowania w postaci ∀*x*¯(*E*′(¯*x*) ≡ *E*(¯*x*)) (lub w notacji Datalog, *E*′(¯*x*) ← *E*(¯*x*)).



**Rys. 1.** Ogólne ramy dla jakości danych kontekstowych

Z reguły będziemy mieć kopię S′ schematu S, z relacyjnymi predykatami , Chodzi o to, że *Ri*′ oznacza wersję jakości *Ri*, i ich rozszerzenia mogą być porównywane. Te idealne predykaty są powiązane ze (rozszerzonymi) kontekstowymi schematami przez mapowania *α*. Jeśli mapowanie dotyczy jakości predykatów P, zazwyczaj używamy predykatów *R*P′*i* i mapowań *α*P aby podkreślić zależność od P. W zasadzie każde z odwzorowań na Rysunku 1 może być dowolnego rodzaju LAV, GAV, GLAV plus dodatkowe założenia dotyczące otwartości/zamkniętości biorących udział instancji lub zwykłej definicji widoku.

Na przykład, *γ* są definicjami widoków predykatów jakościowych w zakresie elementów w C, plus możliwych źródeł zewnętrznych w E. Wszystkie te elementy, tak jak w przypadku integracji danych wirtualnych i systemów zarządzania danymi rówieśniczymi, określa zbiór dopuszczalnych instancji I do kontekstowej instancji C.

W kolejnych rozdziałach możemy rozważyć i zbadać kilka istotnych szczególnych przypadków tych ogólnych ram. W każdym z nich, zwracamy uwagę na: (a) Problem oceny jakości instancji D złożonej z relacji *R*1(*D*)*,...,Rn*(*D*). Ma to związek z analizą, jak różnią się one od idealnych, jakościowych instancji *Ri*. (b) Problem charakteryzowania i uzyskiwania wysokiej jakości odpowiedzi na pytania, które według przewidywań będą odebrane przez instancję D, która jest w trakcie oceny.

Jak widać, te dwa problemy są związane ze sobą. Rzeczywiście, początkowy problem może być postrzegany jako szczególny przypadek tego ostatniego. Możemy przedstawić główne idee już w ogólnych ramach: Biorąc pod uwagę różne schematy, mapowania, a niektóre zmaterializowane relacje (np. *Ri*, *Ek*, niektóre (części) *Cj*) relacyjne predykaty *R*′(*I*) lub *R*′P(*I*) będą miały (możliwe że wirtualne) dopuszczalne rozszerzenia, powiedzmy *Ri*′(*I*). Jakość *Ri*(*D*) jest określana przez jej "odległości" do *Ri*′(*I*), np. przez liczność, |*Ri*(*D*) △ *Ri*′(*I*)|, różnicy symetrycznej. Różne funkcje mogą być rozważane, zwłaszcza jeśli istnieje kilka dopuszczalnych rozszerzenia dla *Ri*′, jak to często ma miejsce w przypadku VDISs, gdzie może się pojawić kilka wirtualnych instancji *legalnych*.

W odniesieniu do odpowiedzi na zapytania jakościowe, jeśli zapytanie Q jest postawione do D, ale oczekuje tylko odpowiedzi jakościowych, zapytanie może być zapisane w za pomocą predykatów *Ri*′ i odpowiadać na podstawie ich rozszerzeń.

Przykład 2. (przykład 1 ciąg dalszy) Rozważmy zapytanie o pacjentach i ich temperatury około południa w dniu wrz/5: Q(*p,v*) : ∃*t*∃*d*(*TempNoon*(*p,v,t,d*) ∧ *d* = Sep/5). Jakość odpowiedzi na to zapytanie postawione w Tabeli 1 powinna być ⟨*Tom Waits*, 38.5⟩, a mianowicie występowanie na dwóch pierwszych atrybutach krotki 1, ale nie krotki 2, ponieważ nie spełniają wymagań jakościowych zgodnie z tabelami kontekstowych 3, 4 i 5. Należy zauważyć, że jeśli zamiast jest to samo zapytanie zadane w tabeli 2, który zawiera tylko dane dotyczące jakości w odniesieniu do wymagań jakościowych, mamy dokładnie taką samą odpowiedź.

Jeżeli istnieje kilka dopuszczalnych instancji, tworzących klasę I, schemat S′ może być instancjonalny w nie, uzyskując instancje *R*′(*I*), z I ∈ I. W konsekwencji, S′(*I*) := {*R*′(*I*) | *R* ∈ S} tworzy instancję do schematu S′. W odpowiedzi dotyczące jakości Q ∈ *L*(S), mogą więc być zdefiniowane jako te, które są pewne:

*QAns**,* dla wszystkich *I* ∈ I}

gdzie Q 'jest otrzymywany z Q poprzez zastąpienie predykatami *Ri* ich kopii, *Ri*′. Pojęcie jakości odpowiedzi może być wykorzystane, aby określić jakość np. D: Dla każdej relacji *R* ∈ S, możemy postawić zapytania *R*(¯*x*) i uzyskać odpowiedzi dotyczące jakości *QAns*(R). Każdy z *QAns*(R) staje się instancją dla predykatu R i może być porównany z *R*(*D*).

# Instancje jako widoki i instancje kontekstowe

W szerokim i powszechnym przypadku ogólnych ram, w uzupełnieniu do schematu S, kontekstowy schemat C, oraz instancja *D* od S, mamy następujące:

(a) W każdej CQP *P* ∈ P zdefiniowany jako koniunkcyjny widok, *P*(¯*x*) ← *γC*(¯*x*), pod względem elementów C (i ewentualnie wbudowanych predykatów). Oznaczamy CP schemat C rozszerzony o schemat P.

(b) Dla każdego predykatów *R* ∈ S, jako kopia tego, bazie ∈ R S kopii tego, *R*′, który jest zdefiniowany jako koniunkcyjny widok schematu CP:



gdzie *φ*C*R*(¯*x*)*,φ*P*R*(¯*x*), są w ich kolei koniunkcjami wzorów cząstkowych z predykatami w odpowiednio C, P. Szczególny przypadek uzyskuje się, gdy w (2) nie ma CQPów w definicji widoku:



Jeśli mamy instancję I dla schematu C, wtedy otrzymujemy obliczone rozszerzenia *R*′(*I*) oraz *R*′P(*I*) przez zastosowanie wzoru (2) lub (3). Teraz, gdy mamy także instancję D z S, która jest przedmiotem oceny jakości, to *R*(*D*) może być porównane z *R*′(*I*) oraz *R*′P(*I*).

Intuicyjnie, każdy CQP może być stosowany do wyrażenia cząstkowych wymagań jakości wymaganej przez konsumenta danych lub pokrywanej przez producenta danych. Z CQP możemy ograniczyć dopuszczalne wartości dla niektórych atrybutów krotek w *I*, tak, że tylko krotki jakościowe znajdą drogę do *D*.

Chociaż CQP można wyeliminować poprzez rozkładanie ich wzorów w Datalog, robimy je tutaj wyraźne z kilku powodów: (a) W celu podkreślenia ich roli jako predykatów przejmujących wymagania jakościowe. (b) Pozwalają one nam na porównanie wymagań dotyczących jakości danych w bardziej konkretny sposób. Na przykład, jest oczywiste, że wymóg jakości "wartości temperatury muszą być mierzone przez termometr doustny lub na podczerwień" jest mniej restrykcyjny niż "wartości temperatury muszą być mierzone przez termometr doustny". (c) Nasze podejście pozwala na rozważenie CQP, które nie są określone tylko w kategoriach osobnego C, ale także w odniesieniu do innych źródeł zewnętrznych, jak pokazano na rysunku 1, to znaczy, według widoku definicji w postaci *P(¯x) ← γC (¯x),γE (¯x)*.

## Najprostszy przypadek

Prosty, ograniczony przypadek ogólnych ram, i z jednego w poprzednim rozdziale, w szczególności, występuje wówczas, gdy instancja D w ramach oceny jest dokładnie zmaterializowanym widokiem instancji kontekstowej za pomocą wzoru w postaci (3). To jest, na każdy *R* ∈ S, zakładamy, że *R*(*D*) = *R*′(*I*). Możemy jednak dodać dodatkowe wymagania jakościowe, otrzymując w ten sposób wystąpienie *R*′P(*I*) za pomocą definicji wzoru w postaci (2). Byłoby to idealnym przykładem predykatu *R* otrzymanego z *I* z użyciem dodatkowych warunków i jakości. W tym przypadku, *R*(*D*) = *R*′(*I*), oraz *D*(*I*) := {*R*′(*I*) | *R* ∈ S} = *D*. Posiadamy również poniższą instancję dla schematu S:

*D*P(*I*) = {*R*P′ (*I*) | *R* ∈ S oraz *R*P′ jest zdefiniowany przez (2)}

Zgodnie z oczekiwaniami, mogą istnieć różnice pomiędzy D i *DP*(*I*). Ten ostatni jest uważany za czystą wersję D. Właściwie, to otrzymuje się *R*′P(*I*) ⊆ *R*′(*I*) = *R*(*D*), dla każdego   
*R* ∈ *S*.

*Przykład 3*. (kontynuacja przykładu 1) Schemat S zawiera predykat bazy danych TempNoon(Pacjent, Wartość, Czas, Data) z instancją w Tabeli 1 w ramach oceny. Schemat kontekstowy C zawiera predykaty *S(Data, Zmiana, Pielęgniarka)*, *T(Pielęgniarka, Typ Danych)* i C(Nazwisko, Rok) wprowadzony wcześniej. Mamy przykłady dla nich, odpowiednio: Tabele 3, 4 i 5. Ponadto C zawiera predykat *M(Pacjent, Jakość, Czas, Data, Urządzenie)*, który rejestruje wartości wszystkich pomiarów wykonanych na pacjentach przez pielęgniarki (np. temperatury, ciśnienia krwi, etc.), wraz z ich czasem, datą, zastosowanym typem termometru (na przykład, termometr, pomiar ciśnienia krwi), i jego instancją w tabeli 6.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Pacjent** | **Wart.** | **Czas** | **Data** | **Urządz.** |
| T. Waits | 38.0 | 12:15 | Sep/6 | Therm. |
| ... | ... | ... | ... | ... |
| T. Waits | 37.6 | 10:50 | Sep/7 | Therm. |
| T. Waits | 120/70 | 11:30 | Sep/7 | BPM |
| T. Waits | 37.9 | 12:15 | Sep/7 | Therm. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Pacjent** | **Wart.** | **Czas** | **Data** | **Urządz.** |
| T. Waits | 37.8 | 11:00 | Sep/5 | Therm. |
| T. Waits | 38.5 | 11:45 | Sep/5 | Therm. |
| T. Waits | 38.2 | 12:10 | Sep/5 | Therm. |
| ... | ... | ... | ... | ... |
| T. Waits | 110/70 | 11:00 | Sep/6 | BPM |
| T. Waits | 38.1 | 11:50 | Sep/6 | Therm. |

Relacja *TempNoon(Patient, Value, Time, Date)* dla każdego pacjenta i dnia zawiera pomiary temperatury około południa, może być rozpatrywana jako zmaterializowany widok instancji w Tabeli 6.

Zgodnie z (3) możemy użyć następującej definicji widoku, która uwzględnia pomiary wykonane pomiędzy 11:30, a 12:30.

(5)

Materializując powyższy widok otrzymamy instancję pokazaną w Tabeli 1.

W celu wyrażenia oczekiwań w stosunku do jakości, wprowadzone zostaną predykaty jakości (CQP). Dzięki nim, będziemy w stanie utworzyć relację zawierającą jedynie krotki spełniające wymagania doktora, np. aby pomiar temperatury wykonywany był przez certyfikowaną pielęgniarkę, z użyciem ustnego termometru. Odpowiednio, P = { *Oral(Instr), Certified(Patient, Date, Time), Valid(Value)* }.

Aby spełnić definicje, musimy wprowadzić najpierw predykat pomocniczy, *Temp(Patient,Date,Time,Nurse,Instr,Type)* który przygotowuje informacje o pomiarze temperatury, użytym instrumencie i imieniu pielęgniarki dokonującej pomiaru. *Temp* łączy się z każdym pomiarem, pielęgniarką i rodzajem termometru w zależności od czasu wykonania pomiaru.

Pierwsze dwa predykaty jakości (CQP) mogą zostać zdefiniowane następująco, z pomocą predykatu pomocniczego:

(6)

(7)

Pierwszy predykat jakości jest spełniony tylko wtedy, gdy użytym typem termometru jest termometr ustny (*tp*=oral). Drugi predykat może zostać wykorzystany do określenia czy pomiar został wykonany przez certyfikowaną pielęgniarkę.

Trzeci predykat zapobiega błędnemu uzupełnieniu wartości poprzez sprawdzanie czy wynik pomiaru mieści się w pewnym założonym możliwym przedziale wartości.

(8)

Z pomocą powyższych trzech predykatów, zgodnie z (2) możemy utworzyć nową relację:

(9)

Powyższa relacja powinna zawierać tylko pomiary spełniające wymagania doktora, co odpowiada instancji w Tabeli 2.

## Jakościowa realizacja zapytań

Zapytania są napisane w języku powiązanym ze schematem S, skierowane do instancji D bazy danych. Dobrymi odpowiedziami na zapytania do D powinny być w zasadzie odpowiedzi na to same zapytanie zadane do instancji D’(I) lub D’p(I) . W konsekwencji oraz jako szczególny przypadek (1), dla zapytania , zbiorem *jakościowych odpowiedzi* na Q będzie:

(10)

Dla zapytań monotonicznych, np. złączeń, zawiera , gdzie ostatnie oznacza zbiór odpowiedzi na Q z D.

Odkąd R(D) są uzyskiwane jako zmaterializowane widoki kontekstowej instancji I, odpowiadanie na zapytania może zostać zrealizowane w oparciu o rozwijanie widoków, tzn. w celu uzyskania jakościowych odpowiedzi, określamy zapytanie które będzie kierowane do czystych relacji R’(I):

Algorytm Rozwijania Widoków: (QUA Quality Unfold Algorithm)

1. Zamień każdy predykat R w Q na odpowiadający mu R’ (lub R’P), uzyskując zapytanie Q’.
2. Zamień zapytanie Q’ na zapytanie , przez rozwijanie oparte na (2).
3. Jeśli trzeba, rozwiń definicje CQP, uzyskując „jakościowe zapytanie” , które może być zrealizowane na I.

Ostatni krok algorytmu otwiera możliwość rozważania CQP które nie są oparte tylko o schemat C i opierają się o zewnętrzne źródła, jak również reprezentujące inne predykaty jakości jak te przedstawione w [19].

Jakość D może być określona przez porównywanie każdego R(D), które wiąże się z R’p(I). Jest to tylko szczególny przypadek odpowiadania na zapytania: Zbiór czystych odpowiedzi na atomowe zapytania może być porównywany z odpowiadającymi R(D).

*Przykład 4.* Rozważmy zapytanie do schematu S naszego bieżącego przykładu, które pyta o temperaturę pacjenta, dnia 5 sierpnia (Sep/5): , które w notacji Datalog wygląda następująco:

(11)

W celu uzyskania jakościowych odpowiedzi, Q jest przepisywane pod względem schematu S’. Q jest trywialnie przepisywane jako: . Teraz TempNoon’P jest zdefiniowane przez (9). Przez rozwijanie widoku uzyskujemy:

(12)

To zapytanie może być wykonane bezpośrednio na I, które zawiera relację M, poprzez rozwijanie definicji (6)-(8) predykatów jakości lub bezpośrednie używanie ich rozszerzeń, jeśli zostały zmaterializowane.

Zauważmy, że w (12) mogliśmy mieć predykaty które nie są zdefiniowane tylko pod względem kontekstowego schematu C, ale również pod względem innych zewnętrznych źródeł. W konsekwencji zapytanie nie może być wykonane tylko na I, i może wymagać dodatkowych danych.

*Przykład 5.* (ciąg dalszy przykładu 4) Rozważmy teraz sytuację, w której zamiast posiadania instancji C(Nurse, Year), mamy jej definicję w znaczeniu zewnętrznego źródła #X(Nurse), które zawiera informacje certyfikowanych pielęgniarkach; #X(Nurse) zwraca *true* jeżeli podana pielęgniarka jest certyfikowana i *false* w przeciwnym przypadku. Odkąd C(Nurse, Year) jest częścią definicji predykatu CQP *Certified* (reguła (7)), realizacja nierozwiniętego zapytania w (12) spowoduje żądanie danych z #X(Nurse) w celu wartościowania predykatu *Certified.*

Niezależność predykatów jakości od kontekstowych danych czy źródeł jest szczególnie interesująca w przypadku, gdy chcemy użyć ich do filtrowania krotek z relacji, powiedzmy R w D. Ta sytuacja może być z łatwością uwzględniona w naszych ramach pracy. Dla predykatów , rozważamy kopię lub *alias,* . Każdy R’ posiada atrybuty i argumenty R, oraz ich dziedziny. Mamy również prostą definicję GAV dla , rozważając R jako *właściwe źródło,* w terminologii wirtualnej integracji danych [12] (popularne w definicjach widoków na pojedynczej instancji). Otrzymujemy kopię R’(D) z R(D) jako część kontekstowej instancji, a kontekstowa instancja staje się .

Teraz, gdy chcemy uzyskać jakościowe odpowiedzi na zapytanie Q skierowane do D (ze schematem S), biorąc pod uwagę predykaty jakości, zastępujemy każdy predykat in Q poprzez złączenie . Dane do realizacji dodatkowej formuły jakości ostaną pobrane z zewnętrznego źródła. Wynikowe zapytanie zostanie zrealizowane na I oraz rozszerzeniach dla predykatów jakości. Gdy brak tych drugich, realizacja zapytania może wywołać zapytania do zewnętrznych danych.

W tej części rozważamy wygodny, lecz niekoniecznie częsty przypadek, gdy instancja D, na której dokonujemy oszacowania jakości, jest kolekcją zmaterializowanych widoków instancji kontekstowej I. Alternatywne i naturalne przypadki, które musimy rozważyć mogą posiadać jedynie częściową kontekstową instancję I- wraz z metadanymi kontekstowych odniesień. Przestudiujemy to w części 4.

# Brakujące kontekstowe dane

W przeciwieństwie do tego co mogło być zasugerowane w powyższych przykładach, nie możemy założyć, ze zawsze mamy dostęp do kontekstowej instancji I dla schematu C. Mogą to być niekompletne dane (puste) instancji I-, jak również dane z zewnętrznych źródeł. W tym bardziej ogólnym przypadku pojawia się sytuacja podobna do tych badanych w ramach systemów wirtualnej integracji danych. Tutaj, kontekstowy schemat zachowuje się jak pośredniczący, globalny schemat, a instancja D jako zmaterializowane źródło danych. W dalszej części będziemy się dokładniej zajmować tym połączeniem.

Załóżmy, że nie mamy kontekstowej instancji I dla schematu C, np. . Możemy zobaczyć D jako źródło danych dla systemu wirtualnej integracji danych , z globalnym schematem rozszerzającym kontekstowy schemat C [22, 4]. Możemy założyć że wszystkie dane w D są powiązane z przez C, jednak może zawierać więcej danych tego samego typu co te wniesione przez D. W konsekwencji, zakładamy że D jest otwartym źródłem dla . To założenie będzie ujęte poniżej w założonych i prawdziwych globalnych instancjach dla .

Jako że nie wszystkie dane w D muszą spełniać wymagania jakościowe w związku z C, potrzebujemy uwierzyć w relację pomiędzy D i jego spodziewaną jakościową wersją. W tym celu, podobnie jak w poprzednich przypadkach rozszerzamy C kopią S’ schematu Teraz i również staje się częścią globalnego schematu .

**Definicja 1.** Załóżmy że każde jest zdefiniowane jako widok Datalog: . (a) Prawdziwą instancją systemu jest instancja I- globalnego schematu, takiego jak: (a1) Dla każdego (a2) . (b) Instancją I schematu C jest prawdziwa kontekstowa instancja (LCI legal contextual instance) jeżeli istnieje I’ dla , taka że (restrykcje I’ dla schematu C).

Warunek (a1) szczególnie pociąga dane z D w kierunku . Prawdziwe instancje mają rozszerzenia danych z D gdy widoki definiujące predykaty R są obliczane. Wyrażenia w (a2) zachowują się jak globalne warunki integralności, np. na schemacie C’ i powodują efekt czyszczenia danych wirtualnie ładowanych do .

Możemy również rozważyć modyfikację tego przypadku gdzie w dodatku do D mamy jedynie fragment I- potencjalnej instancji kontekstowej. Oznacza to że mamy niekompletne dane kontekstowe. W tym przypadku Definicja 1 musi zostać zmodyfikowana poprzez dodanie warunku na I: (a3) , który wymaga aby prawdziwa instancja I’ była „zgodna” z częściową instancją I-. Przy , uzyskujemy poprzedni warunek. (przypis: Pojawienie się częściowej i zmaterializowanej globalnej instancji I- może być ujęte w scenariuszu wirtualnych systemów integracji danych, przez uznawanie I- jako oddzielnego właściwego źródła dla .)

**Definicja 2.** Krotka jest jakościową odpowiedzią na zapytanie jeżeli , gdzie Q’ jest uzyskane z Q poprzez zamianę każdego przez R’.

Jak poprzednio przyjmujemy jako zbiór jakościowych odpowiedzi na zapytanie Q do D przez C.

*Przykład 6.* Przyjrzyjmy się raz jeszcze zapytaniu Q(p, v) w (12) z Przykładu 4. Niech Q’(p,v) oznacza to samo zapytanie, teraz w związku ze schematem S’. Instancja D *TempNoon(Patient, Value, Time, Date)* w Tabeli 1 podlega ocenie jakości.

Teraz definiujemy wirtualny system integracji danych , z D jako otwartym źródłem, i relacjami w tabelach 3,4 i 5 tworzącymi częściową globalną instancję I-. W tym przypadku brakuje instancji (relacji) dla predykatu *M(Patient, Value, Time, Date, Instr)* w C.

Zgodnie z definicją 2, jakościowa odpowiedź na Q(p,v) musi zostać uzyskana z każdego LCI dla . Teraz każda LCI będzie zawierać krotki z *TempNoon(Patient, Value, Time, Date)* spełniające warunki stawiane w (9). W rzeczywistości tabela 2 odnosi się do najmniejszej LCI dla : Żaden jej podzbiór nie jest LCI ani żaden nadzbiór spełniający (9) również nie jest LCI. W konsekwencji pierwsza krotka tabeli 2 jest jedyną spełniającą dodatkowy warunek d = Sep/5. Otrzymujemy:

Odkąd zdefiniowaliśmy początkową instancję D jako otwarte źródło, możemy czerpać korzyści ze wszystkich algorytmów dla obliczania pewnych odpowiedzi na zapytania, które zakładają otwartość źródła [17]. Zakładając że te zapytania i definicje widoków są połączone, możemy wykorzystać np. *algorytm odwróconych zasad* [14]lub jego rozszerzenia [4, 11]. Zilustrujemy jego wykorzystanie w następującym przykładzie.

*Przykład 7.* (kontynuacja przykładu 6) Jeżeli odwrócimy definicję *TempNoon* w (5) otrzymamy:

(13)

Możemy zrealizować zapytanie w (12) poprzez rozwijanie definicji predykatu M, zgodnie z (13), uzyskując:

Przepisane zapytanie może być teraz zrealizowane na instancjach *TempNoon, S, C i T* (odpowiednio tabele 1, 3, 4, i 5). W ten sposób uzyskujemy taki sam wynik jak w poprzednim przykładzie.

Wreszcie, w związku z oceną jakości danych powstało kilka alternatyw. Jeżeli chcemy ocenić D, możemy rozważyć dla każdego LCI I instancję . Mamy . Możliwą miarą jakości może być . Analiza i porównanie tych dwu możliwych miar jakości zostanie dokonana w przyszłych pracach.

# Powiązane prace

Badania nad jakością danych obejmują zarówno charakteryzację typów błędów, modelowanie procesów, w których mogą one powstawać jak również opracowywanie sposobów ich wykrywania oraz eliminacji. Większość podejść do powyższych problemów, bazuje na domniemanym założeniu, że błędy w danych pojawiają się szczególnie na poziomie symboliczno-syntaktycznym, tzn. rozbieżności pomiędzy wartościami (przykładowo Kelvin, a Kelvn).

Jak pokazano w [19], problemy z jakością danych pojawiają się także na poziomie semantycznym, tzn. jako rozbieżności pomiędzy nadawanymi im znaczeniami i interpretacjami. Dokładniej, nawiązując do [19], problemy takie mogą pojawić się w przypadku różnicy między zamierzonym znaczeniem (przypisanym przez producenta danych) i interpretacją ich znaczenia (dokonaną przez konsumenta danych). Taka rozbieżność często spowodowana jest niejednoznacznością w komunikacji między producentem, a konsumentem danych. Taka wieloznaczność jest nieunikniona, gdy zmienne czynniki (jak typ użytego termometru lub stan pacjenta) nie są wyraźnie odnotowane w danych (bądź metadanych). Oczywiście to, czy taka wieloznaczność jest uznawana za problem z jakością danych, zależy od celu w jakim dane te są używane.

W [19] przedstawiono zarys metod definiowania zarówno syntaktycznej, jak i semantycznej jakości danych w jednolity sposób, opartych na podstawowych pojęciach wartości i znaczeń. Zostało również przedstawionych wiele wysokopoziomowych predykatów jakości danych, opartych na porównywaniu symboli i znaczeń (ścisłe dopasowanie, częściowe lub jego brak). W niniejszej pracy podejmujemy kroki, mające na celu zaproponowanie konkretnego mechanizmu zdobywania i porównywania na poziomie semantycznym wymagań dotyczących jakości danych, korzystając z kontekstowych relacji i predykatów jakości oraz pokazania ich zastosowań w realizacji zapytań.

Użycie kontekstów w zarządzaniu danymi było proponowane już wcześniej (w [9] znajdują się badania), jednakże wykorzystywało inne sposoby zdobywania, reprezentowania oraz wykorzystania kontekstów. Przykładowo, kontekstowe informacje były wykorzystywane do wspierania półautomatycznych procesów projektowania widoków (opisane w [8]). Kontekst w [8] składa się z wielu elementów będących parami klucz-wartość (np. stanowisko=”przedstawiciel”, sytuacja=”na miejscu”, czas=”dziś”). Pewne graniczenia również mogą być przedstawiane w kontekście (przykładowo gdy stanowisko to „menadżer”, sytuacja nie może odbywać się „na miejscu”).

Określenie kontekstu pozwala na wybór małego widoku spośród dużej struktury bazy danych, który jest uznawany za trafny w tym kontekście. Projektowanie widoku będącego świadomym kontekstu, gdy kontekst jest doprecyzowany może odbywać się ręcznie lub półautomatycznie poprzez łączenie częściowych widoków zależnych od poszczególnych elementów kontekstu [10]. W niniejszej pracy kontekst jest określony w podobny sposób, jednak w innym celu niż w [8]. Głównym celem [8] jest redukcja rozmiarów (np.: wydzielenie użytecznych danych w danym kontekście), natomiast niniejsza praca jako główny cel stawia wybór danych w zależności od ich jakości (np.: znalezienie takiego podzbioru danych, który najlepiej spełnia pewne wymagania jakościowe).

Pracą najbardziej powiązaną z szacowaniem jakości danych jest [24]. Propozycja Naumann’a bazuje na uniwersalnej relacji [23] utworzonej z globalnego, relacyjnego schematu do integracji niezależnych źródeł danych. Zapytania są zbiorem atrybutów uniwersalnej relacji, z możliwymi warunkami nałożonymi na wartości atrybutów. Aby odwzorować zapytanie w źródłowych widokach, zapytania użytkownika są tłumaczone na zapytania do globalnego schematu bazy danych. Naumann wyróżnia kilka kryteriów jakości źródeł danych, takich jak ich wiarygodność, obiektywność, reputacja oraz weryfikowalność. Kryteria te, są wykorzystywane przy określaniu modelu jakości dla zapytań.

Zgodnie z [24], jakość zapytania jest określana następująco. Każde źródło otrzymuje punktację jakości informacji (IQ – information quality) dla każdego kryterium uznawanego za istotne. Punktacje te są następnie formowane w wektor, gdzie każdy element odpowiada innemu kryterium. Użytkownicy mogą wyrazić swoje preferencje dotyczące danego kryterium poprzez nadawanie wag poszczególnym elementom wektora, tworząc wektor ważony. Wektor ten jest wykorzystywany z kolei przez wieloatrybutowe metody podejmowania decyzji (MADM multi-attribute decision-making) w celu utworzenia rankingu źródeł danych wchodzących w skład uniwersalnej relacji. Metodami tymi może być zarówno proste skalowanie i sumowanie punktacji jak i złożone wzory oparte o macierze zgodności. Model jakości nie zależy od wybranych metod MADM, jeżeli korzysta z wag użytkownika i punktacji jakości IQ. Celem jest, biorąc pod uwagę wektory IQ źródeł, uzyskanie wektora IQ planu zapytania zawierającego źródła. Plan zapytania może być rozumiany jako drzewo połączeń między źródłami: liśćmi są źródła danych, natomiast pozostałymi węzłami są złączenia. Punktacje jakości informacji są obliczane dla każdego wewnętrznego węzła, idąc od liści ku korzeniowi. Ogólna jakość planu zapytania dana jest w postaci punktacji jakości korzenia tego drzewa.

Społeczność zajmująca się reprezentacją wiedzy (ang. knowledge representation community) wykonała pracę nad formalizacją i wykorzystaniem kontekstów. Postały ogólne idee w dziedzinie integracji i współpracy modeli i teorii. W [15] położono nacisk na właściwe interakcje lokalnych środowisk. W nowszych pracach, pojęcie kontekstu, czy też multi-kontekstu, zostało sformalizowane za pomocą *reguł mostowych* pomiędzy oznaczonymi kontekstami, gdzie każdy z nich może posiadać własną bazę wiedzy lub ontologię [12]. Reguły mostowe są wyrażane za pomocą rachunku zdań. Nie jest jasne czy potrafią one oddać bogate zależności znajdujące się w aplikacjach zarządzających danymi. Istnieje jeszcze niedawna praca o integracji rozproszonych logik opisowych [18], niekoniecznie odwołująca się do kontekstów, która pokazuje pomysły podobne do tych z literatury na temat kontekstów i reprezentacji wiedzy.

6. Podsumowanie

Zaproponowaliśmy ramy pracy pozwalające opisać bazę danych pod kątem jej jakości. Czynniki wpływające na jakość zostały ujęte jako interakcje instancji bazy danych z dodatkowymi, kontekstowymi źródłami danych bądź metadanych. W skład naszych ram pracy wchodzi mapowanie źródeł danych, które występuje na przykład w wirtualnej wymianie danych. Dobrej jakości odpowiedzi na zapytania, są zależne od współzależności pomiędzy danymi właściwymi, kontekstowymi i metadanymi.

Są to pierwsze kroki w kierunku wyrażania jakości danych i jakościowego realizowania zapytań jako czynności zależnych od kontekstu. Przeanalizowaliśmy kilka przypadków ogólnego postępowania, jak również dokonaliśmy kilku stwierdzeń na temat mapowania, widoków oraz zapytań. Ogólne ramy postępowania pozostają jednak wciąż do zbadania (więcej w [5, 6]). W tym celu powinno powstać i zostać zbadanych więcej algorytmów pozwalających na szacowanie jakości i dobrej jakości realizację zapytań.

Spośród wielu ciekawych obiektów prowadzonych badań, warto wspomnieć o: (a) Wykorzystaniu danych i zewnętrznych predykatów jakości w opisie instancji bazy danych. (b) Użyciu i integracji bardziej rzeczywistych i bezwzględnych predykatów jakości wprowadzonych w [19], w naszych ramach pracy. Mogą one obejmować tak szczegółowe aspekty jak składnia wartości, poprawność, sens, znaczenie etc. (c) Szczegółowej i porównawczej analizie miar jakości wspomnianych w tej pracy.

Konteksty pojawiły się już w literaturze dotyczącej zarządzania danymi, głównie w związku z oczywistymi aspektami danych jak np. geografia lub wymiary. Naszym zdaniem znacznie brakowało dotąd ogólnego pojęcia kontekstu, jego formalizacji i zastosowania w zarządzaniu danymi. Jest to najważniejszy problem, który musi zostać całościowo zbadany. Zaproponowaliśmy tylko kilka pierwszych pomysłów w tym kierunku.

Źródła

1. Abiteboul, S., Hull, R. and Vianu, V. Foundations of Databases. Addison-Wesley,

1995.

2. Ballou, D., Wang, R., Pazer, H. and Tayi, G. Modeling Information Manufacturing

Systems to Determine Information Product Quality. Management Science,

44(4):462-484, 1998.

3. Batini, C. and Scannapieco, M. Data Quality: Concepts, Methodologies and Tech-

niques. Springer, 2006.

4. Bertossi, L. and Bravo, L. Consistent Query Answers in Virtual Data Integration

Systems. In Inconsistency Tolerance, Springer LNCS 3300, 2004, pp. 42-83.

5. Bertossi, L. and Bravo, L. Query Answering in Peer-to-Peer Data Exchange Systems.

In ’Current Trends in Database Technology’, Springer LNCS 3268, 2004, pp.

478-485.

6. Bertossi, L. and Bravo, L. The Semantics of Consistency and Trust in Peer Data

Exchange Systems. Proc. International Conference on Logic for Programming, Ar-tificial Intelligence, and Reasoning (LPAR 07), 2007, Springer LNCS 4790, pp.

107-122.

7. Bleiholder, J. and Naumann, F. Data Fusion. ACM Computing Surveys, 41(1):1-41,

2008.

8. Bolchini, C., Curino, C., Orsi, G., Quintarelli, E., Rossato, R., Schreiber, F. and

Tanca, L. And What Can Context Do for Data? Communications of the ACM,

52(11):136-140, 2009.

9. Bolchini, C., Curino, C., Quintarelli, E., Schreiber, F. and Tanca, L. A Data-

Oriented Survey of Context Models. SIGMOD Record, 36(4):19-26, 2007.

10. Bolchini, C., Quintarelli, E. and Rossato, R. Relational Data Tailoring Through

View Composition. Proc. ER’07, Springer LNCS 4801, 2007, pp. 149-164.

11. Bravo, L. and Bertossi, L. Logic Programs for Consistently Querying Data Integration

Systems. Proc. International Joint Conference on Articial Intelligence(IJCAI'03), 2003, Morgan Kaufmann, pp. 10-15.

12. Brewka, G. and Eiter, Th. Equilibria in Heterogeneous Nonmonotonic Multi-

Context Systems. Proc. AAAI 2007, pp. 385-390.

13. De Giacomo, G., Lembo, D., Lenzerini, M. and Rosati, R. On Reconciling Data

Exchange, Data Integration, and Peer Data Management. Proc. PODS 2007. pp.

133-142.

14. Duschka, O., Genesereth, M. and Levy, A. Recursive Query Plans for Data Integration.

Journal of Logic Programming, 2000, 43(1):49-73.

15. Giunchiglia, F. and Serafini, L. Multilanguage Hierarchical Logics. Articial Intel-

ligence, 1994, 65:29-70.

16. Grahne, G. and Mendelzon, A. O. Tableau Techniques for Querying Information

Sources through Global Schemas. Proc. ICDT 1999. pp. 332-347.

17. Halevy, A. Answering Queries Using Views: A Survey. VLDB Journal, 2001,

10(4):270-294.

18. Homola, M. and Serafini, L. Towards Formal Comparison of Ontology Linking,

Mapping and Importing. In Proc. DL’10, CEUR-WS 573, 2010, pp. 291-302.

19. Jiang, L., Borgida, A. and Mylopoulos, J. Towards a Compositional Semantic

Account of Data Quality Attributes. Proc. ER’08, Springer LNCS 5231, 2008, pp.

55-68.

20. Kivinen, J. and Mannila, H. Approximate Inference of Functional Dependencies

from Relations. Theoretical Computer Science, 1995, 149:129-149.

21. Kolaitis, Ph. Schema Mappings, Data Exchange, and Metadata Management.

Proc. PODS’05, 2005, pp. 61-75.

22. Lenzerini, M. Data Integration: A Theoretical Perspective. Proc. PODS’02, 2002,

pp. 233-246.

23. Maier, D., Ullman, J. and Vardi, M. On the Foundations of the Universal Relation

Model. ACM Transactions on Database Systems, 1984, 9(2):283-308.

24. Naumann, F. Quality-Driven Query Answering for Integrated Information Systems.

Springer, 2002.

25. Wang, R. and Strong, D. Beyond Accuracy: What Data Quality Means to Data

Consumers. J. Management and Information Systems, 1996, 12(4):5-33.