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 kontektu

# 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.