5. 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”).