A requirement-oriented data quality model and framework of a food composition database system

DISS. ETH NO. 20770

A Requirement-Oriented Data Quality Model and Framework of a Food Composition Database System

A dissertation submitted to

ETH ZURICH

for the degree of

Doctor of Sciences

presented by

KARL PHILIPPE PRESSER

Dipl. Informatik-Ing. ETH

December 3, 1976

citizen of Zurich, ZH, Switzerland

accepted on the recommendation of

Prof. Dr. M. Norrie, examiner
Prof. Dr. H. Hinterberger, co-examiner
Prof. Dr. C. Chan, co-examiner

2012

Abstract

Data quality is an important issue for data acquisition, maintenance and retrieval in an information system because data users demand to obtain best quality data. Data quality research is a relatively young field and is dominated by management scientists who propose solutions for classification and organisational process improvement. Concepts concerning the implementation in information systems are hardly available. For this reason, every information system architect and programmer has to design and implement their own solutions to deal with data quality problems.

To address these problems, we propose a definition of data quality, a data quality model and a data quality framework for IT professionals and users as a contribution to the interdisciplinary field of data quality. As data quality is influenced by multiple user requirements, we put data quality requirements centre stage and named our model and framework Requirement-Oriented Data Quality model (RODQ model) and Requirement-Oriented Data Quality framework (RODQ framework).

The objective of the RODQ model is to help information system architects and users to specify data quality requirements, measure and aggregate them to a total value of data quality. Total quality values can be used to gain an overview of data's quality and to make statements about the quality of considered data. The RODQ model has an RODQ schema with which data quality requirements can graphically be modelled. The RODQ model also includes a classification of data quality requirements according to characteristics that influence their implementation in an information system and includes a simplified assessment schema that can be applied to every data quality requirement.

The RODQ framework is a conceptual framework that proposes implementation concepts to get a comprehensive data quality framework in an information system. To the RODQ framework belongs the concept of data quality prevention with the main idea to evaluate data against RODQ schemas before they are stored in the database and hence to prevent data of low quality entering the information system. This idea is not only designed for graphical user interfaces but also for automated machine-to-machine interfaces. For this reason, the concept of data quality controller is defined as the single quality management place where data quality requirements are defined and evaluated. A further concept is the data quality analysis, which has the objective to support data quality sustainability and which is similar to data quality prevention. The difference is that data quality analysis operates on existing data in the database and in addition is able to analyse sets of data records. The idea behind this concept is also to present quality information to users while they are working with the system. Consequently, not only the data maintainer can use data quality information but also every data user. An additional aim of data quality prevention and data quality analysis is to make the data maintainer and data user quality conscious for the data by providing support at different places in an information system. Furthermore, the long-term analysis of data and their quality is a specific requirement for

scientific data. In addition to the existing concept of data versioning, we present the concept of data quality versioning that performs data quality measurements periodically. The resulting amount of quality values can become problematic over time and we also propose a solution to manage this amount of values.

The showcase application is our implementation of a food composition database management system, called FoodCASE, in which food items and nutrient values are managed by food scientists. Our main quality focus is therefore on empirical and scientific data collections. We finally present the implementation of all concepts in FoodCASE and provide a retrospective and critical discussion about our concepts.

Zusammenfassung

Datenqualität ist ein wichtiger Aspekt bei der Erfassung, Wartung und Abfrage von Daten in einem Informationssystem, weil Datennutzer die beste Qualität erwarten. Das Forschungsgebiet der Datenqualität ist ein relativ junges Gebiet und wird von Wirtschaftswissenschaftlern dominiert, die Lösungen für die Klassifizierungen von Qualitätsaspekten und organisatorische Prozessverbesserungen vorschlagen. Konzepte zur Umsetzung dieser Lösungen in Informationssysteme sind kaum vorhanden. Aus diesem Grund muss jeder Architekt und Programmierer von einem Informationssystem das Datenqualitätsmanagement selber konzipieren und realisieren.

Um diese Probleme anzugehen, definieren wir den Begriff der Datenqualität auf eine neue Weise. Wir schlagen auch ein Datenqualitätsmodell und ein Datenqualitätsframework vor, das sowohl den Entwickler als auch den Benutzern unterstützen soll. Weil Datenqualität durch mehrere Benutzeranforderungen beeinflusst wird, stellen wir diese Anforderungen in den Mittelpunkt und benennen unser Modell und Framework Requirement-Oriented Data Quality-Modell (RODQ Model) und Requirement-Oriented Data Quality Framework (RODQ Framework).

Das Ziel des RODQ Model ist es, Architekten und Anwendern von Informationssystem zu helfen, Anforderungen an die Datenqualität zu formulieren, diese zu messen und zu einem Gesamtwert an Datenqualität zusammenzufassen. Die zusammengefassten Qualitätswerte können verwendet werden, um einen Überblick über die Qualität von Daten zu gewinnen und um Qualitätsaussagen über die betrachteten Daten machen zu können. Das RODQ Model verfügt über ein RODQ Schema mit denen Anforderungen an die Datenqualität grafisch modelliert werden. Das RODQ Model enthält auch eine Klassifizierung von Qualitätsanforderungen nach Merkmalen, die die Umsetzung in einem Informationssystem beeinflussen. Das RODQ Schema beinhaltet auch ein vereinfachtes Bewertungsschema, das auf jede Datenqualitätsanforderung angewendet werden kann.

Das RODQ Framework ist ein konzeptionelles Framework, das Implementierungskonzepte enthält um ein möglichst umfassendes Qualitätshandling in einem Informationssystem zu erhalten. Zum RODQ Framework gehört das Konzept der Prävention mit der Hauptidee neue Daten gegen das RODQ Schema zu prüfen, und zwar bevor sie in der Datenbank gespeichert werden umso die Eingabe von Daten minderer Qualität zu vermeiden. Diese Idee ist nicht nur für grafische Benutzeroberflächen, sondern auch für automatisierte Maschine-zu-Maschine-Schnittstellen zweckmässig. Aus diesem Grund wird der Begriff des Datenqualität-Controllers eingeführt der einen zentrale Einheit beschreibt, wo Qualitätsanforderungen an Daten definiert und ausgewertet werden. Ein weiteres Konzept ist die Datenqualitätsanalyse, die das Ziel hat die Nachhaltigkeit der Datenqualität zu unterstützen und Ähnlichkeiten zur Prävention aufweist. Der Unterschied ist, dass die Datenqualitätsanalyse auf vorhandenen Daten in der Datenbank arbeitet und zusätzlich in der Lage ist, mehrere Datensätzen zusammen zu analysieren. Eine weitere Idee hinter diesem Konzept ist die Mitteilung von Qualitätsinformation für den Benutzer

während er im System arbeitet. Folglich können nicht nur die Personen Qualitätsinformationen sehen, die die Daten pflegen sondern auch Personen, die die Daten nutzen. Ein weiteres Ziel der Prävention und der Datenqualitätsanalyse ist es, das Qualitätsbewusstsein von Benutzern zu stärken und zu unterstützen indem an verschiedenen Orten in einem Informationssystem Qualitätsinformationen angezeigt werden. Darüber hinaus ist die Langzeitanalyse von Daten und deren Qualität eine spezifische Anforderung für wissenschaftliche Daten. Zusätzlich zum bestehenden Konzept der Versionierung von Daten, präsentieren wir das Konzept der Versionierung von Datenqualität, die Datenqualitätsmessungen in regelmäßigen Abständen durchführt und abspeichert. Die hieraus resultierende Menge an Qualitätswerten kann in seinem Umfang im Laufe der Zeit problematisch werden und wir schlagen eine Lösung vor wie diese Menge an Werten verwaltet werden können.

Die Implementierung von den oben erwähnten Konzepten haben wir im Informationssystem FoodCASE durchgeführt, mit dem Lebensmittel und deren Nährstoffzusammensetzungen verwaltet werden können. Der Schwerpunkt unserer Untersuchungen lag daher auf empirischen und wissenschaftlichen Datensammlungen. Am Ende dieser Arbeit präsentieren wir die Umsetzungsdetails aller Konzepte in FoodCASE und diskutieren unsere Konzepte retrospektiv und kritisch.