

DISS. ETH Nr. 12391

**Produktdaten-Design und Daten-Integration
durch Zusammenfügen
autonomer Schema-Module**

ABHANDLUNG
Zur Erlangung des Titels
DOKTOR DER TECHNISCHEN WISSENSCHAFTEN

der
EIDGENÖSSISCHEN TECHNISCHEN HOCHSCHULE
ZÜRICH

vorgelegt von

Felix Jakob Metzger
Diplom-Physiker Universität Dortmund
geboren am 25. September 1957
aus Deutschland

Angenommen auf Antrag von:
Prof. Dr. Max Engeli, Referent
Prof. Dr. Moira Norrie, Korreferentin

1997

Abstract

To communicate information related to industrial products — a task achieved by exchanging product data between computer systems — is nowadays an important precondition for a company to be successful. To ensure the transfer of data containing high-level semantic information, without loss of information, generally acknowledged conventions are needed such as those contained in the international standard STEP.

The formal language EXPRESS used for the description of data schemata, is a key element for the STEP standard. Implementations of EXPRESS software exist and, additionally, work is being done to create commercial processors for the exchange of CAD data, which are defined in the STEP Application Protocol (AP) 203. Further AP projects are being standardized.

STEP is influenced by three data models which serve as its conceptual basis: the Relational, the Entity-Relationship and the Object-Function data model. Graphical schema modelling is also an important tool. The STEP architecture is presented and analysed, including organizational aspects which relate to the data integration strategy. STEP is based upon constructs (entities) which are specified in EXPRESS. These AP entities are derived from the Integrated Resources (IRs), a series of parts of the standard which is common to all APs. The meaning of the AP entities — except those entities in the CAD area — is defined mainly by a correspondence with an AP specific reference model. To use several APs in a collaborative way, where different APs describe different aspects, is a problematic task.

The EXPRESS language is based upon entities with formal consistency conditions for the data. Entity data types are defined by the declaration of entities, where the combination of declared entities is possible. By a proposed set of definitions, it is possible to capture the EXPRESS type system precisely in an axiomatic way. Object identity, which is a property of instances of entity data types, represents the distinction of objects in real world. It is shown that the set of all instances for a single entity data type cannot be defined in a satisfactory way. Instead, the set of all instantiations of a whole data schema exists. It is proposed to extend the EXPRESS language by binding algorithms to data structures, and by mechanisms to formally control the use of constructs within another schema.

The STEP standard contains the definition of an interactive database call interface, which is being implemented within European research projects. STEP also contains a file interface, which is the basis for commercial software for the exchange of CAD data. During the implementation of an EXPRESS compiler, which is mapping the information from a data schema into an intermediate code to be executed by a run-time system, concepts have been developed for the generation of storage layouts guaranteeing fast access to the data. This work of the author focused on the creation of layouts for complex entity data types, which are generated by a combination of declared entities. Such layouts are created on demand at run-time, because there exist a lot of such combinations. Combining formally unrelated entities from different schemata is proposed. This way complex entity data types are defined, which can be utilized as a mechanism for the collaborative use of several modularized data schemata describing different aspects.

It is shown that the definition of the meaning of data is done mainly by describing entities as semantic concepts within a schema using descriptive text written in general language, because formalism is not a powerful enough tool to capture real world meaning. It is analysed how the jargons used for different technical disciplines create a problem. STEP intends to build a centrally integrated data schema, formed by all APs collectively, where the integration of jargons is also intended to form a uniform product data language. As a mechanism to achieve this result, AP entities are derived as being specializations of entities from the IRs, where those IR entities often show only structural properties and can hardly be considered as semantic concepts. By modelling through formal consistency conditions which describe entity specializations, and by describing the semantics of AP entities through their correspondence with constructs from the reference model, the documentation of the standards tends to become huge and difficult to read. Even the integration of jargons is not achieved in the end, because the IR entities are only structural concepts, to which hardly any meaning is assigned with respect to the real world!

An alternative is proposed, where the goal of jargon integration is given up. Separated data schema modules are defined, which describe aspects of products within the boundaries of a technical discipline, and which should be ready for a collective implementation without any further specialization. Those modules contain formally unrelated, autonomous data schemata. To connect such data schemata — across the boundaries of modules — the instantiation of schema-spanning complex entity data types is proposed. Entities are combined with each other, each of which being described within one module as a semantic concept, where the module's jargon shall be used. This way, connecting autonomous data schema modules is shown to be possible, connecting can be done independently from the formal structure of the schemata, it can be done after the standardization process has been completed, in order to equip a given technical system with communication software.

Zusammenfassung

Die Kommunikation von Information im Zusammenhang mit industriellen Produkten, die als Austausch von Produktdaten zwischen Computer-Systemen erreicht wird, ist heute eine wesentliche Voraussetzung für eine erfolgreiche Unternehmung. Um die Übertragung von Daten, die hochwertige semantische Information enthalten, verlustfrei zu gewährleisten, bedarf es allgemeingültiger Festlegungen, wie sie in der internationalen Norm STEP enthalten sind.

Die formale Sprache EXPRESS zur Beschreibung von Datenschemata ist eine wesentliche Basis der STEP-Norm. Implementationen befassen sich mit EXPRESS-Software, daneben wird kommerziell an Prozessoren zum Austausch von CAD-Daten gearbeitet, die im Application Protocol (AP) 203 definiert sind. Weitere AP-Projekte befinden sich im Prozess der Normung.

STEP ist beeinflusst durch drei Datenmodelle, die als konzeptionelle Grundlagen dienen: dem Relationalen, dem Entity-Relationship und dem Object-Function Datenmodell. Die graphische Schema-Modellierung ist ein wesentliches Werkzeug. Die STEP-Architektur wird dargelegt und analysiert, inklusive organisatorischer Aspekte, die sich auf die Daten-Integrations-Strategie beziehen. STEP beruht auf in EXPRESS spezifizierten Konstrukten (Entities), die abgeleitet sind aus den Integrated Resources (IRs), einer Serie von Teilen der Norm, welche allen APs gemeinsam ist. Die Bedeutung der Entities im AP wird — abgesehen vom CAD-Bereich — im wesentlichen durch eine Korrespondenz mit einem AP-spezifischen Referenz-Modell festgelegt. Die gemeinsame Verwendung von mehreren APs, die verschiedene Aspekte beschreiben, hat sich als problematisch erwiesen.

Die EXPRESS-Sprache beruht auf Entities mit formalen Konsistenzbedingungen für die Daten. Mit der Deklaration der Entities werden Entity-Datentypen definiert, die auch auf der Kombination von Entities beruhen können. Mit einer Reihe von vorgeschlagenen Definitionen ist es möglich, dieses Datentyp-System axiomatisch präzise zu erfassen. Objekt-Identität als Eigenschaft der Ausprägungen von Entity-Datentypen bildet die Unterscheidung von Objekten in der realen Welt ab. Es wird gezeigt, dass die Menge aller Ausprägungen eines einzelnen Entity-Datentyps nicht befriedigend definiert werden kann. Statt dessen existiert die Menge aller Ausprägungen für ein gesamtes Datenschema. Die Erweiterung der EXPRESS-Sprache wird vorgeschlagen, und zwar durch das Anbinden von Algorithmen an Datenstrukturen und durch Mechanismen zur formalen Kontrolle der Verwendung von Konstrukten in anderen Schemata.

Die STEP-Norm enthält die Definition einer interaktiven Datenbank-Schnittstelle, die in europäischen Forschungsprojekten implementiert wird, neben der File-Schnittstelle, die die Basis darstellt für kommerzielle Software zum Austausch von CAD-Daten. Im Rahmen der Implementation eines EXPRESS-Compilers durch den Autor, der die Schema-Information in einen Zwischencode zur Ausführung durch ein Laufzeit-System abbildet, sind Konzepte entwickelt worden zur Erzeugung von Speicher-Layouts, mit denen der schnelle Zugriff auf die Daten sichergestellt ist. Dabei ist das Erzeugen von Layouts für die komplexen Datentypen wesentlich, die durch Kombination von Entities entstehen. Solche Layouts werden erst zur Laufzeit bei Bedarf erzeugt, da sehr viele Kombinationen existieren. Die Kombination von

Entities, die in keiner formalen Beziehung zueinander stehen und aus verschiedenen Schemata stammen, wird vorgeschlagen. Dadurch werden komplexe Datentypen definiert, die als Mechanismus zur gemeinsamen Verwendung von modularisierten, verschiedene Aspekte beschreibende Datenschemata verwendbar sind.

Es wird dargelegt, dass die Definition der Bedeutung von Daten im wesentlichen geschieht, indem die Entities als semantische Konzepte im Schema mit beschreibenden, alltagssprachlichen Texten versehen werden, weil Formalismen nicht mächtig genug sind zum Erfassen von Bedeutung bezogen auf die reale Welt. Es wird analysiert, welche Probleme dabei die in verschiedenen technischen Disziplinen verwendeten Jargons darstellen. STEP strebt an, dass alle APs zusammen ein zentrales, integriertes Datenschema bilden, womit auch die Integration der Jargons zu einer einheitlichen Produktdaten-Sprache bezweckt wird. Als Mechanismus dient das Ableiten der AP-Entities als Spezialisierungen von Entities aus den IRs, wobei letztere häufig nur strukturellen Charakter haben und kaum noch als semantische Konzepte angesehen werden können. Durch die Modellierung mit formalen Konsistenzbedingungen zur Beschreibung der Spezialisierung eines Entitys, und durch die Beschreibung der Semantik der Entities im AP durch die Korrespondenz mit den Konstrukten im Referenz-Modell, wird die Dokumentation der Normen umfangreich und schwer lesbar. Auch die Integration der Jargons gelingt ungenügend, da die IR-Entities lediglich strukturelle Konzepte sind, denen kaum eine Bedeutung im Sinne der realen Welt zukommt.

Eine Alternative zur STEP-Architektur wird vorgeschlagen, wobei auf die Integration der technischen Jargons verzichtet wird. Dabei werden zur Beschreibung von verschiedenen Aspekten der Produkte getrennte Datenschema-Module definiert, die entlang der Grenzen der technischen Disziplinen gegeneinander abgegrenzt sind. Ohne Spezialisierung sollen die Konstrukte in den zusammen verwendeten Datenschema-Modulen für die Implementation geeignet sein. Die Module enthalten formal nicht verknüpfte, autonome Datenschemata. Es wird vorgeschlagen, die Verbindung der Schemata miteinander — über die Grenzen der Datenschema-Module hinweg — durch Ausprägen von Schema-übergreifenden, komplexen Datentypen herzustellen. Dabei werden Entities miteinander kombiniert, die innerhalb je eines Moduls mit dem dort verwendeten Jargon als semantische Konzepte beschrieben werden sollen. Es wird gezeigt, dass auf diese Weise das Zusammenfügen von autonomen Datenschema-Module möglich ist, dass es unabhängig von der formalen Struktur der Schemata nach dem Abschluss des Prozesses der Normung vorgenommen werden kann, um ein gegebenes technische System mit Kommunikations-Software auszustatten.