Metadatenbasierte Konzepte für die Migration und Integration von Daten

Cristian Pérez de Laborda
cristian@pdl.de

Lehr- und Forschungseinheit für Datenbanksysteme
Institut für Informatik, Ludwig-Maximilians-Universität München
Betreuung der Diplomarbeit: Prof. Dr. Stefan Conrad*

Seit der Einführung der elektronischen Datenverarbeitung müssen die Bestände verschiedener Datenbanksysteme miteinander synchronisiert werden. Dazu werden zwischen den beteiligten Komponenten Daten ausgetauscht. Dies geschieht in der Regel durch sogenannte *Flatfiles* – Textdateien in einem vereinbarten Format. Heute gibt es Datenbanken, die auf den Austausch von Daten mit heterogenen Datenquellen spezialisiert sind. In der Praxis haben sich Data-Warehouses und Metadirectories durchgesetzt.

Die für das Management relevanten Informationen liegen in den meisten Unternehmen auf verschiedenen Datenbanken verteilt. Deshalb entstanden Ende der 80er Jahre Data-Warehouses aus der Notwendigkeit heraus, Mitarbeitern die managementrelevanten Informationen des gesamten Unternehmens über einen zentralen Datenbankserver anzubieten [DM88]. Die Daten werden dazu aus meist heterogenen Datenbanken extrahiert und in die integrierte Datenbank des Data-Warehouses importiert [ZGMHW95]. Diese Daten sind dadurch über einen *Single Point of Access* abrufbar. Die Besonderheit des Data-Warehouses ist, historische Daten nicht zu ersetzen, sondern zu archivieren. Damit verfügt man über eine langfristig persistente Datenbasis, mit deren Hilfe Management-Entscheidungen getroffen und später auch nachvollzogen werden können [Inm96, S. 33ff].

Entsprechend den objektrelationalen Datenbanken wurde das Konzept der Objektorientierung auch auf hierarchische Datenbanken angewendet. Die so entstandenen X.500 Directories [Int01a] wurden zunächst hauptsächlich für die Benutzerverwaltung von Unternehmensnetzwerken eingesetzt. Damit erhöhte sich in den meisten Unternehmen die Anzahl der Datenbanken, in denen parallel Personendaten gepflegt werden mussten, auf mindestens zwei: Das Directory und die Personaldatenbank. Das Konzept des Metadirectorys wurde entwickelt, um die Datenpflege zu erleichtern. In das Metadirectory werden unter anderem Personendaten automatisch aus einer beliebigen Datenquelle importiert und von dort aus an andere Komponenten weitergeleitet. Es fungiert dabei nicht nur als Datenvermittler zwischen verschiedenen Datenbanken, sondern bildet ein *Single Point of Administration* für die über das Metadirectory synchronisierten Daten [Int01b], [Enc02].

Sowohl beim Data-Warehouse als auch beim Metadirectory kann der Datenimport folgendermaßen schematisiert werden: Veränderte Daten der Quelldatenbanken müssen erkannt, extrahiert, transformiert und in die eigene Datenbank importiert werden. Darüberhinaus erlaubt ein Metadirectory auch das Verändern der Daten durch den Benutzer in der eigenen Datenbank. Diese modifizierten Daten werden dann an andere Komponenten weitergeleitet. Wie man sieht, spielt der Datenaustausch bei beiden Systemen eine zentrale Rolle.

*Praktische Informatik/Datenbank- und Informationssysteme, Heinrich-Heine-Universität Düsseldorf

Die Diplomarbeit behandelt die grundsätzlichen Problemfelder, die bei dem Zusammenspiel von Daten und Metadaten während einer Datenübertragung auftreten. Dazu werden zunächst alle relevanten Datenflüsse von Data-Warehouse-Systemen und Metadirectories analysiert. In durchaus komplexen Systemen wie Data-Warehouses wird mehr als ein Datenaustausch eingerichtet, um alle internen Komponenten mit den aktuellen Daten zu versorgen [BG01, S. 36]. Aus folgenden Gründen muss darauf geachtet werden, dass alle Datenaustauschbeziehungen eines Systems in einem einheitlichen Format durchgeführt werden:

- erhöhte Übersichtlichkeit aller Datenflüsse eines Systems,
- Möglichkeit auf Standards zurückzugreifen - keine Entwicklung eigener Formate nötig,
- vereinfachte Ankopplung externer Komponenten,
- vereinfachte Administration und Fehlersuche,
- neue Datenflüsse können mit wenig Aufwand hinzugefügt werden.

Bei einem Datenaustausch müssen neben den eigentlichen Daten auch Metadaten, also Daten über Daten, übertragen werden. Darunter versteht man bei Datenbanken nicht die vom Dateninhalt abgeleitete Information, wie zum Beispiel ein Schlagwort, sondern jene, die vom Schema oder von der Datenbank selbst abhängig ist. Dazu gehören Informationen über die

- Basisrelationen,
- Attribute,
- Sichten,
- Indexe,
- Speicherorganisation,
- Benutzer,
- Zugriffsrechte,
- Schlüssel,
- Konsistenzbedingungen und
- Datentypen.

Diese üblicherweise im Repository aufbewahrten Metadaten [HL93, S. 237] werden in der Arbeit daraufhin untersucht, ob eine Übertragung zu einem Austauschpartner sinnvoll ist oder nicht. Die Übertragung von Schemainformationen wie Primärschlüssel wird sich nämlich eher rechtfertigen lassen, als die der Speicherorganisation oder der Indexe.
Zum Abschluss wird eine Aufstellung der heute gängigen Metadatenstandards, wie beispielsweise Dublin Core, MPEG-7, SyncML oder RDF gegeben. Daraufhin wird verdeutlicht, dass sich mit dem *Resource Description Framework* (RDF) nicht nur Webseiten für das *Semantic Web* aufbereiten lassen (vgl. [BLHL01]), sondern auch die in dieser Arbeit erarbeiteten relevanten Metadaten einer relationalen Datenbank repräsentieren lassen. Damit wird RDF zum idealen Datenaustauschformat für Metadaten einer Datenbank. Diese sind mit Hilfe von RDF von Maschinen nicht nur lesbar, sondern auch interpretierbar.

Ausgewählte Literatur

[BG01] Bauer, Andreas (Hrsg.) ; Günzel, Holger (Hrsg.): *Data- Warehouse- Systeme. Architektur, Entwicklung, Anwendung.* Heidelberg : dpunkt, 2001

[BLHL01] Berners-Lee, Tim ; Hendler, James ; Lassila, Ora: The Semantic Web. In: *Scientific American* (2001), Mai

[DM88] Devlin, Barry A. ; Murphy, Paul T.: An architecture for a business and information system. In: *IBM Systems Journal* 27 (1988), Nr. 1, S. 60–80

[Enc02] Enck, John. *2H02 Metadirectory Service Market Magic Quadrant.* Gardner Research Note. 2002

[HL93] Habermann, Hans-Joachim ; Leymann, Frank: *Repository: eine Einführung.* München : Oldenbourg, 1993

[Inm96] Inmon, William H.: *Building the Data Warehouse.* 2. New York : John Wiley & Sons, 1996

[Int01a] International Telecommunication Union. *The Directory: Overview of concepts, models and services. ITU-T Recommendation X.500.* 2001

[Int01b] Internet Applications Group: *DC-Directory and DC-MetaLink Product Overview.* v 2.0. Enfield: Data Connection Limited, 2001

[ZGMHW95] Zhuge, Yue ; García-Molina, Héctor ; Hammer, Joachim ; Widom, Jennifer: View maintenance in a warehousing environment. In: *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, ACM Press, 1995, S. 316–327