



THINKING
NETWORKS

DB-Prism: Integrierte Data Warehouses und Knowledge Networks für das Bank Controlling



DB-Prism: Integrierte Data Warehouses und Knowledge Networks für das Bank Controlling

Übersicht

DB-Prism ist ein integriertes Data Warehouse System, das bei der Deutschen Bank für ein verteiltes Finanz- und Management Controlling (Datenerfassung, Verarbeitung und Reporting) entwickelt wurde. Das System vereinigt die detailgetreue Verfügbarkeit von historischen Daten mit hochaktuellen Reporting- und Planungsmöglichkeiten. Zu den signifikanten Komponenten gehören ein OLAP-System im Terabyte-Bereich und eine Metadatenbank; diese ist zuständig für den gesamten Prozess – von der Auswahl heterogener Daten bis hin zu individuellen OLAP Report-Positionen und zurück zur individuellen Geschäftsaktivität mit Drill-through. Für diese Zwecke wurden sowohl auf der Nutzerebene als auch auf der Metadatenebene Control-Workbenches entwickelt.

1

Einleitung

Das Controlling Warehouse System DB-Prism ist ein wissensgesteuertes Netz von Datenbank-, Verarbeitungs- und Reporting-Teilsystemen. Das volle Funktionsspektrum dieses Systems wird gegenwärtig beim Controlling der Deutsche Bank AG, der Online Bank DB 24 AG und weiteren deutschen Tochtergesellschaften einschließlich DB Trust, DB Lübeck usw. verwendet. Das System wird in der Finanzbuchhaltung, z.B. zum Erstellen von Bilanzen und Gewinn/Verlust-Aufstellungen, zur Information der Bankaufsichtsbehörden und für die täglichen internen Status-Reports verwendet.

Die Genehmigung zum gebührenfreien Kopieren des gesamten vorliegenden Materials oder Teilen dieses Materials wird unter der Voraussetzung erteilt, dass die Kopien nicht zur Erlangung eines direkten kommerziellen Nutzens hergestellt oder verteilt werden, dass das VLDB Copyright vermerkt wird, dass Titel und Datum der Veröffentlichung auftreten und der Vermerk beigefügt wird, dass eine Kopiergenehmigung der VLDB-Stiftung (Very Large Data Base Endowment) vorliegt. Andernfalls ist für die Kopien bzw. für eine Wiederver-

öffentlichung eine Gebühr und/oder eine spezielle Genehmigung der Stiftung erforderlich.

Beim entscheidungsorientierten Rechnungswesen wird das System z.B. zur Bewertung von Wirtschaftsunternehmen, Geschäftsbereichen sowie zur Absatzauswertung im kommerziellen Banksektor verwendet.

Das DB-Prism-Projekt war durch Probleme angeregt, die im Zusammenhang mit Datenqualitätsfaktoren [WSF] auftreten, zum Beispiel bei der Integrität und Transparenz/Nachweisbarkeit von Daten. Zu einer heterogenen Hardware-/Software-Landschaft kamen unterschiedliche Modellierungs- und Aggregationsprinzipien für mehrdimensionale Finanzdaten hinzu, die auf die Geschäftsziele individueller Geschäftseinheiten ausgerichtet waren oder sich einfach aufgrund der Geschichte ergaben.

Die Regelsysteme waren aber nicht nur heterogen, sondern oft auch im Programmcode versteckt. Darüber hinaus durchliefen die Daten häufig unterschiedliche Teile verschiedener Controlling-Systeme, bevor sie in einem Report erfasst wurden. Diese semantische Heterogenität und die in den individuellen Transformationsprogrammen und im globalen Controlling Workflow verborgenen Wissensstrukturen hatten zur Folge, dass die Analysten viel Zeit darauf verwendeten, scheinbar widersprüchliche Report-Informationen miteinander in Einklang zu bringen.

Das primäre Designvorhaben von DB-Prism bestand deswegen darin, diese Heterogenität auf der konzeptuellen, logischen und physikalischen Ebene mit dem Ziel herauszuarbeiten, die Datenintegrität zu erhöhen, die Zusammenhänge zwischen den verschiedenen Datenquellen und den Reporting-Perspektiven explizit zu machen (und damit die Transparenz und Nachweisbarkeit zu verbessern) und dafür zu sorgen, dass Reporting und Auffrischen von Daten hocheffiziente Vorgänge sind. Dieses Ziel wurde einerseits dadurch erreicht, dass kompliziertes Bereichs- und Systemwissen über das Informationsnetz in einer zentralen Metadatenbank erfasst wird und ande-



rerseits ein effizient skalierbares mehrdimensionales OLAP-System verwendet und maßgeschneidert gestaltet wird. Im Ergebnis wird die Heterogenität jetzt in einer einheitlichen Data Warehouse-Umgebung konsolidiert, die auf einem vereinheitlichten Speicher historischer Geschäftsdaten basiert.

2 Die Architektur von DB-Prism

Das System wurde auf eine effiziente Verarbeitung von Massendaten und ein schnelles Reporting abgestimmt. Die Datenbank wird täglich aktualisiert, wobei die laufenden Veränderungen ebenso berücksichtigt werden, wie die kundenorientierten Teildatenwürfel und Reports. Detailgetreue historische Daten werden in allen Einzelheiten vier Jahre lang aufbewahrt.

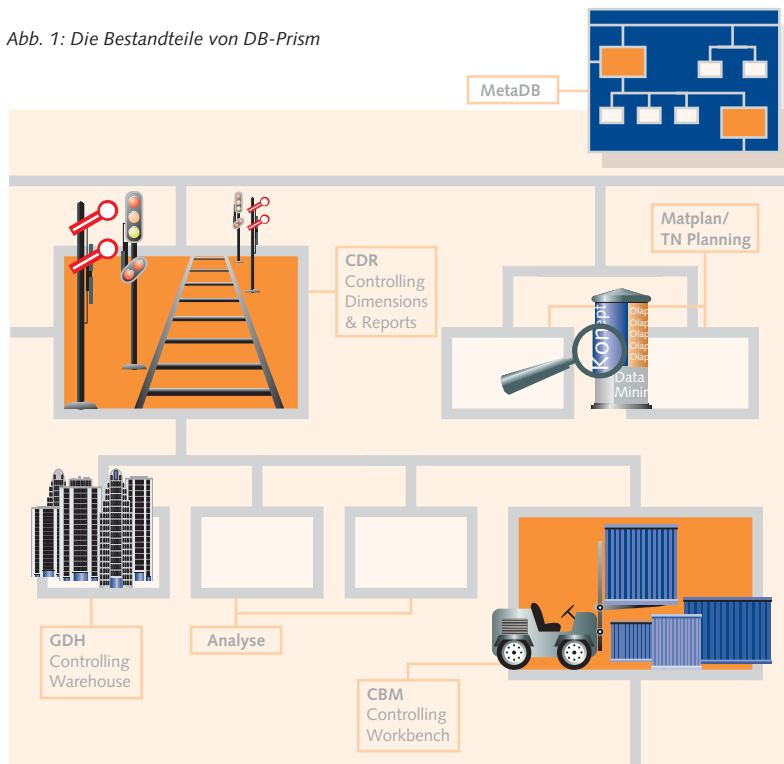
Das Warehouse umfasst 1200 GB, von denen 250 GB die in zweidimensionalen Dateien gespeicherten Basisdaten des

Controlling Warehouse (GDH) ausmachen, während 150 GB aus DB2-Relationen bestehen; man beachte, dass diese voraussichtlich weiter wachsenden Größen die relativ kurze Betriebsgeschichte des Systems widerspiegeln. Der größte Teil und das wichtigste Analyseinstrument umfassen 800 GB verketteter Datenwürfel, die als Gesamtheiten von schwach-besetzten Matrizen in einem mehrdimensionalen OLAP-System gespeichert sind, das den Namen Matplan/TN Planning [TN] trägt.

Im nachfolgenden Teil dieser Arbeit beschreiben wir die in Abbildung 1 dargestellten individuellen Komponenten der DB-Prism-Architektur. Im unteren Teil dieser Abbildung sehen wir einerseits das fundamentale Controlling Warehouse (GDH), das einen gemeinsamen bereinigten Basisspeicher bereitstellt, der sich aus heterogenen Datenquellen zusammensetzt; andererseits ist dort die Controlling Workbench CBM dargestellt, die es dem Controller ermöglicht, mit GDH zu arbeiten. Für weitere Einzelheiten verweisen wir auf Abschnitt 3. In der Mitte

der Abbildung sehen wir auf der rechten Seite – angedeutet durch den Analysten – die OLAP-Umgebung Matplan/TN Planning, welche Datenwürfel enthält, die auf der Grundlage der verschiedenen Kundeninteressen aus dem GDH extrahiert und reorganisiert wurden; Einzelheiten werden in Abschnitt 4 beschrieben. Der Komplex GDH/CBM mit den zugrundeliegenden (und nicht dargestellten) Datenquellen einerseits und die OLAP-basierte Analyse andererseits sind auf eine metadaten-gesteuerte Weise miteinander verbunden. Die in der CDR-Workbench für Controlling-Dimensionen und Reports definierten und entwickelten Metadaten werden in einer Metadatenbank MetaDB gespeichert. Dieser Vorgang wird in Abschnitt 5 beschrieben. Verwendet man die Begriffe der traditionellen Data Warehouse Architektur, wie sie z.B. in [JLVV] beschrieben wird, dann entspricht

Abb. 1: Die Bestandteile von DB-Prism



der Komplex GDH/CBM ungefähr einem integrierten, bereinigten und historisierten, aber noch nicht aggregierten operationalen Datenspeicher, während Matplan/TN Planning einem MOLAP-Warehouse und Kundenumgebung entspricht. Der Komplex MetaDB/CDR stellt eine signifikante Anreicherung der in anderen Data Warehouses verfügbaren Metadaten-Möglichkeiten dar, was hauptsächlich auf die semantische Reichhaltigkeit und Heterogenität auf den konzeptuellen Ebenen sowie der Ebenen der logischen Datenmodelle, der physikalischen Effizienz und der Sicherheit zurückzuführen ist.

3

Das fundamentale Controlling Warehouse

Wie bereits erwähnt, handelt es sich bei den Daten aus heterogenen Quellen um Daten, die in einem Controlling Warehouse (GDH) bereinigt, integriert und historisiert werden, wobei das Warehouse von der CBM Controlling Workbench unterstützt wird. Diese unterstützt nicht nur generische Versionen der obengenannten Operationen, sondern auch bereichsspezifische Operationen für das Finanz- und Management-Controlling.

3.1 GDH (Controlling Warehouse)

GDH basiert auf detaillierten Informationen auf Account-Ebene und ist auf die Bedürfnisse der Massendatenspeicherung ausgerichtet. GDH hat fünf Bestandteile: Import, Transformation, Speicherung, Zugriffsmethoden und Export. Hierbei kommen die folgenden Schlüsselideen zum Tragen:

- *Standardschnittstellen für Import und Export.* Die Daten werden aus vielen heterogenen Betriebssystemen, statistischen Archiven und Controlling-Anwendungen hauptsächlich über Standardschnittstellen-Dateien importiert. Das Liefersystem benötigt kein spezielles GDH-Know-how, da die Dateien ausschließlich durch GDH-Module erstellt werden. Zum Aufrufen dieser Module sind standardisierte Dateischnittstellen verfügbar. Diese werden durch die MetaDB erzeugt und spezifisch an das individuelle Liefer-

system angepasst. Darüber hinaus findet der Datenexport nur über Standardschnittstellen statt, die durch den GDH-Zugriff bereitgestellt werden.

- *Transformation und "Bereinigung" von Importdaten* führen zu einer konsistenten Datenbank mit *vereinheitlichter Datenbedeutung*. Der Transformationsvorgang beinhaltet Restrukturierung, Neudefinition, Filterung, Ableitung, Extrapolation und Aggregation von Daten und Datenregistern. Schlüssel und Datenfelder aus heterogenen Liefersystemen werden harmonisiert, aktuelle und historische Daten unterliegen der gleichen Art von Strukturprozessen.

- *Homogene Speicherung aktueller und früherer Daten; Speichermethoden, die mit Zugriffsanforderungen synchronisiert sind.*

Es gibt eine Vielzahl von Zugriffsanforderungen, die auf dem GDH plaziert sind: Einfügung von Massendaten, Austausch und Löschen ganzer Datengruppen, Zugriff auf individuelle Abschlüsse und Datengruppen, Verarbeitung sämtlicher individuellen Abschlüsse. Derart vielfältige Zugriffsanforderungen verlangen unterschiedliche Speichermethoden, falls die Datenverfügbarkeit und die Antwortzeiten akzeptabel bleiben sollen. Dieser Umstand erfordert die Akzeptanz von Redundanzen bei der Datenspeicherung. Die Redundanzen erweisen sich als nützlich, solange deren Integrität innerhalb des Systems gewährleistet ist.

- *Datenzugriff nur über "lizenzierte" Zugangsmodule.* Das GDH erscheint externen Nutzern als Black Box. Der Datenzugriff wird über Zugriffsfunktionen bereitgestellt. Die Verkapselung des GDH bewirkt eine strikte Trennung der logischen und physikalischen Aspekte. Dieser Umstand ermöglicht eine flexible Datenverwaltung und Datenspeicherung hinsichtlich der Erweiterbarkeit von Strukturen oder Veränderungen in der physikalischen Datenbank.

Weitere wichtige Features sind:

- **die Möglichkeit flexibler und effizienter Wiederholungen.** Die GDH-Bestandteile für Import, Transformation, Speicherung und Export lassen sich – sogar retroaktiv – effizient und flexibel wiederholen. Eine Wiederholung wird hinunter bis zur kleinsten Lieferbestandsebene und auf allen betroffenen Warehouse-Ebenen ausgeführt. Wiederholungsanforderungen werden reorganisiert und automatisch verarbeitet.
- **Integriertes Workflow- und Kontrollmanagement.** Die GDH-Programme und Module arbeiten universell über ein gemeinsames Steuer- und Kontrollzentrum. Eine universelle MetaDB zeichnet alle Verarbeitungsschritte und Statusauswertungen durch nachfolgende Programme auf. Auf diese Weise erfolgt eine ereignis- und ergebnisorientierte Verarbeitung.
- **Skalierbarkeit.** Die Strukturen im GDH sind flexibel angelegt und gestatten eine nachfolgende Erweiterung (variables Containerkonzept). Strukturelle Änderungen lassen sich daher in allen relevanten Bestandteilen relativ schnell und kostengünstig implementieren. Es ist kein Problem, neue GDH-Datenbestände in die GDH-Verwaltung zu integrieren.

3.2 CBM (Controlling Workbench)

CBM ist eine zentrale Workbench für Controller, die mit GDH arbeiten. Die Workbench unterstützt das Hinzufügen von relevanten Informationen, Korrekturen, Zwischenkontoverschiebungen und Buchungen sowie Plausibilitäts-Checks. Aus diesem Grund hat CBM Zugriff auf die individuellen Accounts und Themenfelder des GDH.

4

TN Planning (mehrdimensionales OLAP-System)

Matplan/TN Planning ist eine mehrdimensionale Datenbank und ein OLAP-System in der TN Planning-Umgebung. Matplan/TN Planning wird von TN entwickelt und vermarktet, einer in Aachen ansässigen Software-Firma [TN]. Matplan/TN Planning ermöglicht eine äußerst flexible OLAP-

Funktionalität (z.B. Drill-down und Roll-up zwischen unterschiedlichen Dimensionen und Aggregationsebenen) und liefert verschiedene Anwendungsmodule für Planung, Konsolidierung, Datamining, Datenintegration mit XML und dynamisches Preisgestaltungs- und Erlösmanagement.

Bei DB-Prism wird Matplan/TN Planning als Standardinstrument für Reporting und Analyse verwendet. Es lassen sich interaktive, komplexe und mehrdimensionale Analysen durchführen, da es sich um ein aufgabenorientiertes System für den Endnutzer handelt.

Matplan/TN Planning basiert auf aggregierten Reporting-Daten, die mit Hilfe des Controlling Warehouse (GDH) konstruiert werden.

Zu den Hauptmerkmalen von DB-Prism gehören:

- **die vollständige Integration von Matplan/TN Planning in das Metadatenkonzept.** Sämtliche Matplan/TN Planning Reportdaten, die dem GDH entnommen wurden, sind in der Metadatenbank bekannt. Die Dimensionen, die Quelle, Bedeutung, Struktur und Zusammensetzung dieser Daten werden registriert. Die MetaDB verwaltet alle Reportdefinitionen zur Datenbetrachtung und sämtliche Metadaten, die für die Struktur der mehrdimensionalen Datenwürfel relevant sind.
- **Drill-through Funktionalität.** Der Nutzer ist dazu in der Lage, die Basis der aggregierten Matplan/TN Planning Reportdaten, die individuellen GDH-Abschlüsse als Analysespur aufzurufen. Zum Zweck einer fokussierten Analyse können Auswahlkriterien zur Datenanalyse definiert werden (Dimensionen und ihre charakteristischen Merkmale, Abkürzungen).
- **Skalierbarkeit & Speicherkapazität.** Im Interesse unbegrenzter analytischer Möglichkeiten könnte es als nützlich erscheinen, sämtliche Dimensionen als orthogonal zu deklarieren, d.h. sämtliche Daten so zu definieren, dass sie auf allen Aggregationsebenen durch alle Dimensionen intrinsisch teilbar sind. Das würde jedoch eine Speicherkapazität von $8,56 \times 10^{26}$ Zellen erfordern, wozu alle auf der Welt vorhandenen Computer nicht ausreichen würden.



Aus diesem Grund gibt es mehrere gekoppelte Datenwürfel unterschiedlicher Dimensionen und Skalen. Diese Datenwürfel bilden einen logischen Hyperwürfel, so dass sämtliche Möglichkeiten einer mehrdimensionalen Navigation (z.B. Drilldown/ Roll-up, Slice & Dice) für das globale System vollständig verfügbar bleiben.

- *Skalierbarkeit & Performance.* DB-Prism muss ein sehr großes Volumen aktueller Daten bewegen, damit täglich ein vollständig aufgefrischtes Data Warehouse zur Verfügung steht. Matplan/TN Planning ermöglicht ein flexibles Skalieren, indem man frei zwischen der Prä-Aggregation in Stapelfenster und der dynamischen Online-Aggregation hin- und herschaltet. Eine mögliche Zeitverzögerung, die mit der dynamischen Online-Aggregation zusammenhängt, wird durch die Verwendung effizienter Hashing-Techniken vermieden.

5

Metadaten-Management

Wie bereits erwähnt, stellte die semantische Heterogenität bezüglich der im Programmcode verborgenen Domain-Regeln eine der Hauptmotivationen für das DB-Prism-Projekt dar. Damit war – wie in [JJQV] postuliert – eine die Bedeutung der Daten und deren Beziehungen dokumentierende konzeptuelle Perspektive zu einem kritischen Erfolgsfaktor des Systems geworden. Als gleichermaßen wichtig erwies sich jedoch auch die Abbildung dieser konzeptuellen Annäherung auf die logische Ebene – d.h. die inkrementale Erstellung der grundlegenden GDH-Daten aus verschiedenen Quellen und die Umstrukturierung der GDH-Daten in zahlreiche Würfel-formate und Reportstrukturen.

In Anbetracht der Größe und Komplexität des Systems ist es nicht überraschend, dass auch die Optimierung der physikalischen Ebene sehr wichtig ist. Zu den signifikanteren gelösten Problemen auf der physikalischen Ebene gehören zum Beispiel das Scheduling von Vereinigungsoperationen bei der Datenintegration zur GDH-Produktion sowie die Entscheidung, wann Würfel prärealisiert werden sollen und wann eine dynamische Aggregation durchzuführen ist.

Diese Aspekte sind eng miteinander verknüpft und können

nicht unabhängig voneinander betrachtet werden. Die präzise Dokumentation und – nach Möglichkeit – Automatisierung dieser Aufgaben ist das Ziel der Metadaten-Management-möglichkeiten bei DB-Prism; analog zu GDH/CBM umfassen diese Einrichtungen die Metadatenbank MetaDB und die als CDR bezeichnete zugehörige Workbench.

5.1 MetaDB

Die Metadatenbank von DB-Prism bringt das Wissen über ein breites Spektrum heterogener Anwendungs- und Systemaspekte zusammen. Die MetaDB beschreibt Daten vom Standpunkt ihres Ursprungs, ihrer Struktur, Bedeutung, Nutzung und ihrer Beziehungen; es handelt sich dabei um die Wissensbasis für das gesamte System und für die Wechselwirkung seiner Bestandteile. Als aktives Knowledge Networking System kennt, lenkt und steuert die MetaDB die Prozesse innerhalb des Gesamtsystems nicht nur auf der konzeptuellen Ebene, sondern auch auf der logischen und physikalischen Ebene. Auf diese Weise werden Transparenz, Integrität, Qualität, Standardisierung, Konsistenz und Flexibilität gewährleistet. Abbildung 2 stellt die Schritte der DB-Prism-Operationsprozesse vom Quellenimport via GDH und der OLAP-Würfelerstellung bis hin zum Drill-through und zurück zu den operationalen Daten dar; damit deutet die Abbildung die Reichhaltigkeit des durch die MetaDB abgedeckten Wissens an.

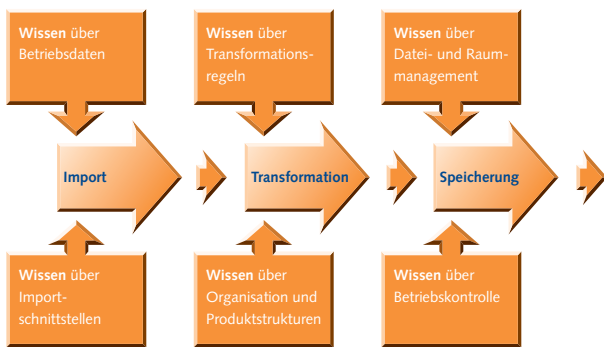
5.2 CDR Tool (Controlling Dimensionen & Reports)

CDR ist das zentrale Benutzerwerkzeug für Controller und Systemexperten zur Verwaltung der MetaDB. Dimensionen, Produkte, Hierarchien, Lieferschnittstellen und Reportpositionen können definiert und angepaßt werden. Die Verarbeitung wird auf der operationalen MetaDB Online ausgeführt. Nach einer von einem integrierten Programm gesteuerten Integration und nach der Durchführung umfassender Qualitätsprüfungen können Veränderungen in Echtzeit realisiert werden.

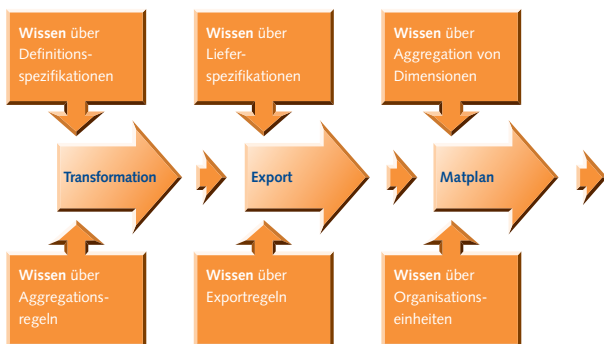
Zusätzliche wichtige CDR-Features sind:

- das **Zeit- und Statuskonzept** zur Definition der Scheduling- und Überwachungsverfahren für operationale Datenflüsse, wie sie in Abbildung 2 dargestellt sind.

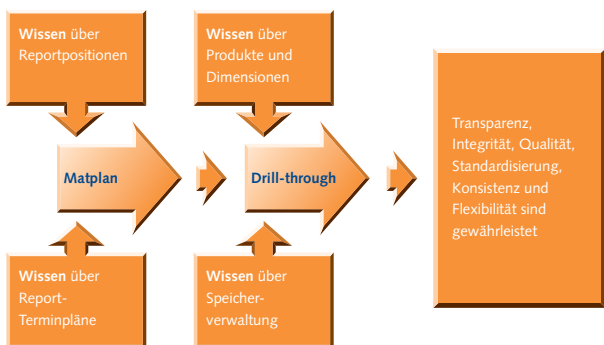




A. GDH Teilprozess Loading/Auffrischen



B. Mehrdimensionaler Aggregations-Teilprozess



C. Mehrdimensionaler Analyse-Teilprozess mit Drill-through

- die **Online-Statuskontrolle** vom Datenimport über die GDH-Erstellung bis hin zum Datenexport
- die **Metadatenexport-Funktionalität** zur automatischen oder interaktiven Strukturentwicklung externer Systeme wie Matplan/TN Planning.

6 Schlussfolgerungen

Das DB-Prism-System ist in der hier beschriebenen Form seit etwa einem Jahr in Betrieb. Es hatte aufgrund folgender Qualitätsmerkmale signifikante Auswirkungen auf die Organisation: Integrität, Kohärenz zwischen Controlling-Pfaden und Reporting-Systemen, Aktualität und Verfügbarkeit von

Abb. 2: Wissensvernetzung im DB-Prism-Prozess

Daten, Entwickelbarkeit und effiziente Handhabung sehr großer Mengen heterogener Daten an der Quelle und kunden-seitig.

Zu den hauptsächlichen Erfolgsfaktoren gehören die von Matplan/TN Planning bereitgestellten hocheffizienten und flexiblen Einrichtungen für mehrdimensionale Datenbanken, ein Metadatenmodell und eine Metadatenumgebung, die bereichsspezifische Qualitätsaspekte aus der konzeptuellen Perspektive des Wissens über das Rechnungswesen von Banken unterstützt. Hinzu kommen traditionellere logische Datentransformationen sowie physikalische Datentransport- und Datenmanipulationsmöglichkeiten. DB-Prism ist also ein interessantes Beispiel einer qualitätsorientierten und konzept-fokussierten Data Warehouse Architektur, wie sie im europäischen DWQ-Projekt untersucht wird [JJQV]. Die laufende Arbeit beinhaltet die Erweiterung auf zusätzliche Bestandteile der Organisation und die Steigerung der Zugriffsmöglichkeiten für die Kunden.

Danksagungen. Die Autoren bedanken sich bei den Kollegen von Thinking Networks in Aachen und Frankfurt, die entscheidend zum Erfolg des hier beschriebenen Systems beigetragen haben.

Autorennachweis:

DB-Prism: Integrierte Data Warehouses und Knowledge Networks für das Bank Controlling
 Elvira Schäfer¹ Jan-Dirk Becker¹ Matthias Jarke²

¹ Global Technologies & Services, Deutsche Bank AG, Prior-Str. 11, 65936 Frankfurt, Deutschland

² GMD-FIT, Schloss Birlinghoven, 53754 Sankt Augustin, Deutschland

Quellenangaben:

[TN] Thinking Networks AG. Matplan and TN Planning product description (in Deutsch). <http://www.thinking-networks.de>, June 2000.

[JJQV] Jarke, M., Jeusfeld, M., Quix, C., Vassiliadis, P. Architecture and quality in data warehouses: an extended repository approach. Special Issue Advanced Information Systems Engineering (Pernici/ Thanos, eds.), *Information Systems* 24(3): 229 - 253, 1999.

[JLVV] M. Jarke, M. Lenzerini, Y. Vassiliou, P. Vassiliadis: *Fundamentals of Data Warehouses*. Springer-Verlag 1999.

[WSF] Wang, R.Y., Storey, V., Firth, C.P. A framework for analysis of data quality research. *IEEE Trans. Knowledge and Data Eng.* 7(4):623-640, 1995.



**THINKING
NETWORKS**

Auszug aus der Referenzliste:

Deutsche Bank

DG Bank

Postbank

SEB Bank

...

Thinking Networks AG

Markt 45 - 47

D-52062 Aachen

Telefon: +49(0)241/47072-0

Fax: +49(0)241/47072-250

info@thinking-networks.com

www.thinking-networks.de

Competence Center in Deutschland:

Aachen, Berlin, Bochum, Frankfurt, Hamburg