Einsatzmöglichkeiten des Semantic Web zur Integration von Data Warehouse und Wissensmanagement am Beispiel von SemTalk®

Christian Fillies, Frauke Weichhardt

Semtation GmbH
Bredower Straße 145
D-14612 Falkensee
cfillies@semtalk.com
fweichhardt@fweichhardt.de

Abstract: Obwohl das Semantic Web primär entwickelt wurde, um den Inhalt von Dokumenten darzustellen, ist es sinnvoll, denselben Ansatz auch auf andere Bereiche anzuwenden, in denen eine gemeinsame Sprache und wohl definierte Begriffe benötigt werden. Dies trifft beispielsweise auf die Inhalte eines Data Warehouse zu. In diesem Beitrag wird ein Projekt aus einer deutschen Krankenkasse vorgestellt, in dem dieses Konzept auf die Einführung und den Betrieb eines Data Warehouse angewendet wurde. Zum Einsatz kam dabei das grafische Modellierungswerkzeug SemTalk®.

1 Semantic Web und Data Warehouse

Das Semantic Web [BHL01] erlaubt es, Informationen nicht mehr nur textuell sondern als „Modell“ formalisiert darzustellen. Es handelt sich dabei um eine Methode, Daten und Metadaten anwendungsunabhängig verteilt zu repräsentieren und für verschiedenartige Anwendungen verfügbar zu machen.[W399],[LS99].

Im Data Warehouse werden Definitionen für Kennzahlen und Kontexte, in denen diese Kennzahlen verwendet werden sollen (Dimensionen), einheitlich bestimmt. Um sie zu verwenden, werden Würfel und Berichte definiert, in denen Kennzahlen und Dimensionen kombiniert werden. Die Dokumentation dieser Metadaten eines Data Warehouse ist von großer Wichtigkeit, da die Anwender wissen müssen, auf welcher Basis ihre Auswertungen durchgeführt werden; das heißt, dass die Definitionen der Elemente nicht nur für die Entwickler sondern auch für die Anwender zur Verfügung stehen müssen. Die Dokumentation der Metadaten lässt sich als Wissensmodell in Form einer Ontologie [Gr95] interpretieren. Sie muss in einer Form durchgeführt werden, die es einerseits dem Entwickler auf einfache und effektive Weise ermöglicht, die von ihm entwickelten

Inhalte darzustellen und andererseits dem Anwender eine einfache und gezielte Form des Zugriffs auf die gewünschten Inhalte zur Verfügung stellt.

Eine grafische Notation der Ontologie mit Hilfe eines Modellierungstools bietet sich hier an, da damit eine effiziente Verwaltung der Metadaten-Dokumentation mit einer einfachen Darstellungsweise kombiniert werden kann. Um die Dokumentation durch verschiedene Entwickler in den einzelnen Fachabteilungen zu ermöglichen, muss die Modellierung losgelöst von einander erfolgen können. Zur Unterstützung dieser Funktionalität bietet sich die Nutzung von Technologien des Semantic Web an, da diese es ermöglichen, die abgebildeten Strukturen lokal zu verwalten und sie trotzdem weiterhin zentral koordinieren zu können.

2 Praxisbeispiel mit SemTalk

Mit der Einführung eines Data Warehouse bei der AOK Berlin wurde die Abbildung der Metadaten des Systems notwendig. Dazu sollten Kennzahlen und Dimensionen sowie ihre Beziehungen und Verwendungen in Berichten und Datenwürfeln dargestellt werden. Hierzu wurde das Werkzeug SemTalk® [FWW02] der Firma Semtation GmbH auf Basis von Microsoft Visio eingesetzt. Es ermöglicht eine Definition der jeweils benötigten Elemente und Attribute des Modells sowie der verwendeten grafischen Elemente.

Jede Definition einer Kennzahl oder einer Dimension wird in diesem Werkzeug als Objekt angelegt, ebenso jeder Würfel und jeder Standardbericht. Für die Objekte können je nach Bedarf Attribute definiert werden. Kennzahlen und Dimensionen werden in ihren jeweiligen Zusammenhängen mit den entsprechend definierten Attributen modelliert (siehe Abbildung 1).

Das verwendete Werkzeug ermöglicht dabei auch eine mehrfache Darstellung desselben Objekts in verschiedenen Kontexten, um ein einfaches assoziatives Suchen zu unterstützen. Auf Basis der definierten Kennzahlen und Dimensionen können Würfel und Berichte dokumentiert werden, indem ihre Inhalte aus diesen Elementen zusammengefügt werden. Diesen spezifischen Kontexten werden die sonstigen Informationen aus den Würfel- bzw. den Berichtsdokumentationen als Attribute hinzugefügt, z. B. Aktualisierungszeitpunkte, Zuständigkeiten für Aktualisierung oder Ansprechpartner für Datenqualität. Der Zugriff für den Anwender wird über eine HTML-Version des Modells realisiert, das im Intranet bereitgestellt werden kann.

Konflikte in der Benennung werden über ein Namensraumkonzept gelöst. Unter einem Namensraum wird dabei ein logischer Bereich verstanden, der sich durch eine einheitliche Begriffswelt definiert. Namensräume können sich überschneiden. Ein Objekt, das unter verschiedenen Namen im Unternehmen verwendet wird (Synonym), erhält dabei zwei verschiedene Namen, die dem jeweiligen Namensraum zugeordnet sind. Auf das Objekt kann mit beiden Namen zugegriffen werden. Für den Fall, dass derselbe Name für verschiedene Objekte verwendet wird, wird für jede Verwendung ein eigenes Objekt erstellt, das dem jeweiligen Namensraum zugeordnet wird. Über die Kombination aus

Namensraum und Name des Objekts ergibt sich eine eindeutige Identifikationsmöglichkeit.

In dem in Abbildung 1 dargestellten Beispiel werden zwei verschiedene Kennzahlen in zwei verschiedenen Bereichen mit demselben Namen "Krankenhausbehandlungsfälle" verwendet:

- die Anzahl der Fälle für das Fallmanagement im Krankenhaus und
- die Anzahl der Fälle für die Verhandlungen mit dem Krankenhaus

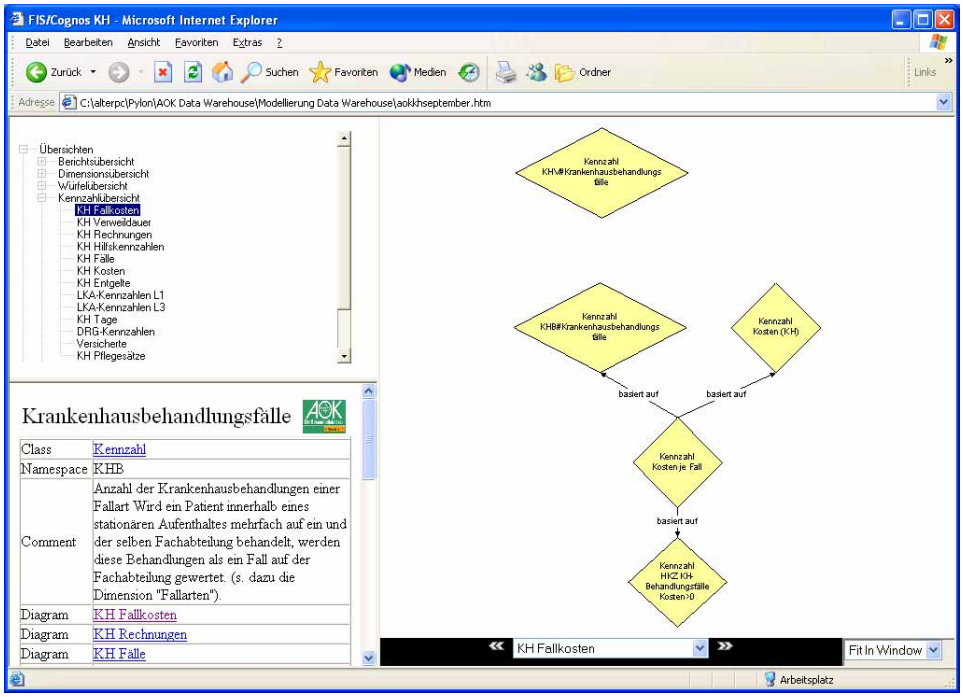

Abbildung 1: Beispiel für die Darstellung des Datenkatalogs im Intranet auf Basis des SemTalk-Modells

Dabei geht es um die Beurteilung der Krankenhäuser bezogen auf die einzelnen Fachabteilungen eines Krankenhauses. Für die Verhandlungen mit dem Krankenhaus ist die Berechnung fachabteilungsbezogener Daten (z.B. Fallzahlen, Verweildauern) gesetzlich geregelt. Vergleichsdaten der AOK Berlin müssen nach den gleichen Methoden ermittelt werden. Interne Verlegungen (Verlegungen von einer Fachabteilung in eine andere innerhalb desselben Krankenhauses) werden in den Berechnungen entsprechend dem Gesetz berücksichtigt. Jede interne Verlegung eröffnet also einen neuen Fall.

Bei der Berechnung fachabteilungsbezogener Daten des Krankenhauses für das Fallmanagement werden interne Verlegungen nur fachabteilungsübergreifend berücksichtigt, da im Fallmanagement der Fall ganzheitlich betrachtet werden muss. Ein neuer Fall entsteht hier nicht, wenn der Patient wieder in die ursprüngliche Abteilung zurückverlegt wird. Nur so lassen sich Ansatzpunkte für das Fallmanagemant ableiten, da bei Betrachtung nur aus der gesetzlich definierten Sicht die durchschnittliche Verweildauer und die durchschnittlichen Kosten je Fachabteilung geschönt werden.

Für das Data Warehouse mussten also zwei verschiedene Kennzahlen mit demselben Namen definiert werden. Über das Namensraumkonzept kann der auftretende Konflikt gelöst werden. Dabei wird für jedes Objekt über den Dateinamen, den Namensraum und den Namen eine URI erzeugt. Auf diese kann speziell bei der verteilten Modellierung auch über das Internet zugegriffen werden (siehe Abschnitt 3). Wie bei allen Semantic Web-Anwendungen wird auf diese Weise sichergestellt, dass alle Beteiligten durch die Benutzung einer solchen URI über dieselbe Sache reden und dass sich Applikationen auf dieselbe Interpretation der Legacydaten beziehen.

3 Verteilung von Modellen

Das Semantic Web eröffnet die Möglichkeit, auf ein zentrales Repository zu verzichten, bei dem alle Beteiligten ihr Wissen in einer konsistenten zentralen Struktur ablegen. Dieses mag zwar beispielsweise für Softwarekomponenten sinnvoll sein, ist aber schon für mittelgroße Unternehmen für Informations- oder Wissensmodelle nicht praktikabel und hemmt entscheidend den einzelnen Mitarbeiter, zum gemeinsamen Modell beizutragen. Ganz praktisch führt ein solcher zentralistischer Ansatz dazu, dass ein großer Teil der Begriffsdefinitionen im Freigabeprozess stecken bleibt und das Gesamtsystem vom Anwender abgelehnt wird. Ebensowenig ist auch die Vorstellung eines zentralen Content Management Systems für das Internet realistisch.

Aus diesem Grund wird hier ein dezentraler Ansatz realisiert, der zentral koordiniert wird. Die Entwickler dokumentieren ihre Modelle lokal. Dabei werden sie durch das Werkzeug auf eventuell bereits existierende Konzepte hingewiesen. Die Modelle werden regelmäßig durch einen Modellkoordinator ausgewertet. Dieser erkennt eventuell notwendigen Diskussionsbedarf und stellt die entsprechenden Begriffe in einem Koordinationsgremium zur Diskussion. Die in dieser Diskussion dann als zentral definierten Begriffe werden in zentralen Modellen als Bausteine abgelegt, die im Anschluss für die Verwendung in den lokalen Modellen zur Verfügung stehen, aber zentral gewartet werden. Über den Einsatz von URI´s (siehe Abschnitt 2) ist der Zugriff auf die benötigten Objekte auch in räumlich verteilten Umgebungen einfach und eindeutig realisierbar.

4 Einsatzperspektiven

Semantic Web-Standards helfen, die Konsistenz in der Verwendung gemeinsamer Begriffe zu sichern. Die Dezentralität des Ansatzes gibt den Fachanwendern die Möglichkeit, eigenes Wissen einfach in ein Netzwerk des Wissens einbringen zu können. Die Verwendung von Semantic Web-Standards zur Wissensmodellierung bietet dem Unternehmen dabei die Möglichkeit, das dokumentierte Domain-Wissen von einzelnen Software-Applikationen (wie einem konkreten Data Warehouse-System) zu separieren. Damit wird die Grundlage geschaffen, das Wissen in anderen Applikationen weiter verwenden zu können.

Das Semantic Web könnte die Architektur von Data Warehouse Anwendungen grundlegend verändern. Es werden zum einen zunehmend Informationen aus Online-Systemen ohne eine spezifische Datenaufbereitung einbezogen. Andererseits wird es eine Mischung aus unternehmensinternen und unternehmensfremden Datenquellen aus dem Internet geben [De02]. Durch das Semantic Web stehen Meta-Informationen über diese Daten und die Web Services, die sie bereitstellen, abgestimmt auf konkrete Geschäftsprozesse zur Verfügung [GM02]. Der vorgestellte Ansatz zeigt, wie sie für den Endbenutzer aufbereitet werden können, damit er leichter verstehen kann, welche externen Daten er mit in sein persönliches Informations- und ggf. Wissenssystem integrieren kann.

Literaturverzeichnis

[BHL01] Berners-Lee, T.; Hendler, J. and Lassila, O.: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American, May 2001.

[De02] Devlin, B.: From Data Warehouse to Information Integration. In: von Maur, E.; Winter, R. (Hrsg.): Vom Data Warehouse zum Corporate Knowledge Center: Proceedings der Data Warehousing 2002. Physica-Verlag, 2002.

[FWW02] Fillies, C.; Wood-Albrecht, G.; Weichhardt, F.: A Pragmatic Application of the Semantic Web Using SemTalk. WWW2002, May 7-11, 2002, Honolulu, Hawaii, USA ACM 1-5811-449-5/02/0005

[GM02] Guha, R.V.; McCool, R.: A System for integrating Web Services into a Global Knowledge Base. http://tap.stanford.edu/ss/ bzw http://www.alpiri.com/sw002.html

[Gr95] Gruber, T.: Towards principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies, 43/1995: S. 907–928.

[LS99] Lassila, O.; Swick., R.: Resource description framework (RDF) - model and syntax specification. Technical report, W3C, 1999. W3C Recommendation. http://www.w3.org/TR/REC-rdf-syntax.

[W399] W3C: RDF Schema Specification. http://www.w3.org/TR/PR-rdf-schema/, 1999.