Semantic Computing und Informationsextraktion als Schlüsseltechnologien für das Produktinformationssystem der Zukunft

Daniel Schuster[1], Matthias Wauer[1], Johannes Meinecke[2], Alexander Schill[1]

[1]Technische Universität Dresden, Fakultät Informatik, Professur Rechnernetze
01062 Dresden
{daniel.schuster, matthias.wauer, alexander.schill}@tu-dresden.de

[2]SAP Research Center Dresden
Chemnitzer Str. 48, 01187 Dresden
johannes.meinecke@sap.com

Abstract. Semantic Computing und Informationsextraktion sind zwei aufstrebende Technologien für das Management von Informationen im Unternehmenskontext. Wir zeigen am Beispiel des Projektes Aletheia wie sich beide Techniken für die Föderation strukturierter und unstrukturierter Produkt-Informationen kombinieren lassen.

Keywords: Semantic Computing, Ontologien, SOA, Informationsextraktion, Informationssysteme, Produktinformationen.

1 Einleitung

Produktinformationen fallen entlang des Produktlebenszyklus in den verschiedensten Formaten und sehr heterogener Granularität an. Noch dazu wirken verschiedene Akteure an dem Entwurf, der Produktion, dem Verkauf und dem Service eines Produktes mit. Die dazu notwendigen Informationen liegen verteilt in verschiedenen Produktinformationssystemen bzw. oft auch in unstrukturierter Form in internen Dateiablagen oder externen Quellen wie Nutzermeinungen im Web vor. Im Projekt Aletheia stellen wir uns der Herausforderung, eine Architektur für ein föderiertes Produktinformationssystem (FPIS) zu entwerfen, das es zu jedem Zeitpunkt des Produktlebenszyklus erlaubt, ein umfassendes Bild der jeweils relevanten Produkt-bezogenen Informationen zu erzeugen. Dazu wird auf Informationen aus Drittsystemen mittels Web-Service-Technologien zugegriffen und schließlich diese Informationen anhand von Domänen-spezifischen Ontologien integriert und für eine einheitliche Informationsdarstellung aufbereitet.

In dem Beitrag stellen wir die beiden aus unserer Sicht dafür notwendigen Schlüsseltechnologien vor, die in Aletheia dazu genutzt werden, der Vision des FPIS näher zu kommen. Durch Techniken aus dem Bereich des Semantic Computing ist es möglich, Domänen-spezifisches Wissen zu modellieren und dieses Wissen zu nutzen, um Informationen aus unterschiedlichen Quellen zusammenzuführen und zu

verknüpfen. Das Gebiet der Informationsextraktion stellt Methoden bereit, um strukturierte Informationen aus unstrukturierten Texten (z.B. im Web) zu gewinnen und somit innerhalb des FPIS nutzbar zu machen. Damit wird nicht nur eine einheitliche Sicht auf existierende Informationssysteme geschaffen, sondern darüber hinaus eine Vielzahl bislang nicht nutzbarer Informationen erschlossen.

2 Föderierte Produktinformationssysteme (FPIS)

Die Idee, Produktinformationen an einer zentralen Stelle innerhalb der Organisation zu verwalten und den Zugriff darauf zu vereinheitlichen ist unter den Begriffen Product Information Management (PIM) bzw. Master Data Management (MDM) von Produktinformationen bekannt. Es gibt dafür eine ganze Reihe von flexibel anpassbaren Lösungen in einem stark wachsenden Markt [9]. Die klassische Architektur eines MDM für Produktinformationen besteht aus einer zentralen Datenbank, die die Produktinformationen und deren Metainformationen verwaltet, sowie den angeschlossenen Systemen (ERP, PLM, lokale Datenbanken), die über eine Messaging-Infrastruktur mit dem MDM-System kommunizieren. Daten werden meist im XML-Format ausgetauscht.

Föderierte Produktinformationssysteme (FPIS) [7] ergänzen diesen Ansatz um die Einbindung unstrukturierter Quellen wie lokaler Dateien (Office-Dokumente, PDF, Scans, E-Mails), Informationen aus dem Web 2.0 (interne und externe Webseiten, Produktforen, Wikis, Blogs), sowie Informationen aus dem Internet of Things (RFID, Sensornetze, …). Solche unstrukturierten Informationen machen nach Schätzungen [5] ca. 50 – 80% der relevanten Informationen im Unternehmen aus.

Dabei erhebt das FPIS jedoch nicht den Anspruch, die wichtigsten Produktdaten zentral vorzuhalten und zu pflegen, sondern erleichtert lediglich den Zugriff auf eine Vielzahl von Informationen, die im Unternehmen an den unterschiedlichsten Stellen vorliegen. Im Sinne einer föderierten Architektur kann ein bestehendes MDM für Produktinformationen an das FPIS angebunden werden und liefert beispielsweise Referenzdaten, die durch aus unstrukturierten Quellen extrahierte Informationen ergänzt werden.

Abb. 1 zeigt die Elemente eines FPIS: Informationsquellen werden über spezielle Extraktionskomponenten eingebunden, die auf die Rohdaten aus den Quellen zugreifen und unstrukturierte Informationen in strukturierte Informationen umwandeln. Ein zentraler Dienst legt diese Informationen in einem Repository ab und erfasst Metainformationen, um so die Herkunft der Daten nachvollziehbar und somit bewertbar zu halten. Ein zentrales Element des FPIS ist das semantische Modell, das Konzepte/Begriffe und deren Beziehungen innerhalb der Organisation definiert und damit die Zusammenführung strukturierter und unstrukturierter Informationen ermöglicht. Der Nutzer bekommt schließlich von einem Application Server eine Oberfläche präsentiert, in der er im gesamten Informationsbestand suchen und navigieren kann.

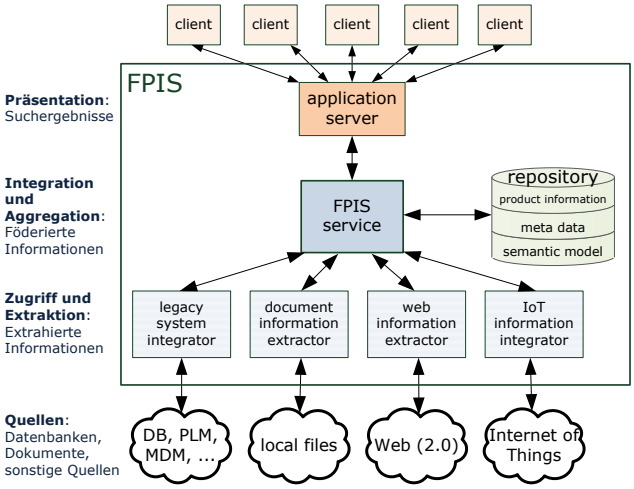

Abb. 1. Föderiertes Produktinformationssystem, nach [7]

3 Semantic Computing und Informationsextraktion in Aletheia

Das Projekt Aletheia [6] ist ein vom Bundesministerium für Bildung und Forschung gefördertes Projekt mit dem Ziel, der Vision eines FPIS näher zu kommen. Fünf Industriepartner (ABB, BMW, DPDHL, Otto Group, SAP) bringen Anwendungsszenarien aus verschiedenen Phasen des Produktlebenszyklus und jeweils unterschiedliche Landschaften von zu integrierenden Produktinformationsquellen ein. Die grundlegende Idee von Aletheia ist es, gemäß der oben dargestellten Vision des FPIS die in den jeweiligen Szenarien benötigten Informationen an einem Zugriffspunkt einheitlich verfügbar zu machen.
Die dabei erhobenen Anforderungen sowie die im Konsens mit allen Partnern entwickelte Referenzarchitektur sind in [8] beschrieben. Techniken der Informationsextraktion werden genutzt, um Datenquellen gemäß der Vision des FPIS (siehe Abb.1) anzubinden. Dabei sind bislang Interfaces für unstrukturierte Dateien auf File Shares, Webseiten, semantische Stores (RDF), Datenbanken sowie spezialisierte Tools zum Export aus Dritt-Anwendungen (z.B. Excel) vorhanden.
Zentrales Element der Föderation ist das Repository, das aus einem semantischen Store zur Speicherung von Domänen-spezifischen Konzepten, Relationen und Fakten, sowie aus einem Store zur Speicherung unsicherer Informationen besteht. Diese Trennung sicherer und unsicherer Informationen ist ein wesentliches Element, da auch der beste automatische Extraktionsprozess nie perfekt arbeitet und somit extrahierte Informationen immer mit Unsicherheit verbunden sind. Deshalb werden in Aletheia Meta-Informationen über Herkunft, Frische und Verlässlichkeit zu extrahierten Informationen gespeichert, die später eine Bewertung der Informationen bzw. deren gesonderte Behandlung bei der Präsentation ermöglichen.

Im Folgenden soll ein kurzer Einblick in die konkret verwendeten Technologien gegeben werden:

Semantic Computing
In Aletheia werden verschiedene Anwendungsontologien in der Sprache F-Logic mit der Software OntoStudio [4] modelliert. Diese umfassen Klassen (Objekte) und deren Relationen untereinander. Des Weiteren werden auch Synonyme und Homonyme sowie Regeln für das Erzeugen neuer Relationen auf der Basis impliziter Relationen erzeugt. In Abb. 2 ist links der F-Logic-Quellcode eines Ausschnitts einer Anwendungsontologie sowie die grafische Darstellung zu sehen. Rechts ist ein Beispiel für eine Regel dargestellt, die eine neue Relation „hasExpert_IR" erzeugt, wenn ein Servicetechniker einen Servicejob für eine bestimmte Maschine ausgeführt hat. In diesem Fall soll der Techniker für konkrete Angebote der Maschine als Experte angezeigt werden.

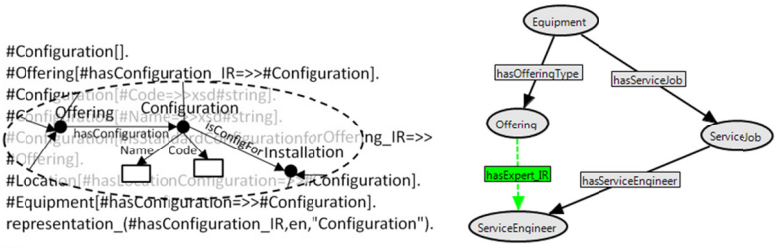

Abb. 2. Beispiele für Anwendungsontologie

Der Zugriff auf das semantische Repository erfolgt mittels der Software OntoBroker [3]. Für die Verwaltung der unsicheren Informationen wird Sesame [2] eingesetzt. Die aus unstrukturierten Quellen extrahierten Informationen werden als RDF-Tripel abgelegt. Der Zugriff erfolgt durch die semantische Anfragesprache SPARQL.

Informationsextraktion
Für die Analyse unstrukturierter Dokumente (Dateien + Web) kommt in Aletheia das Aperture-Framework [1] in Verbindung mit einer Pipeline für die Erkennung von benannten Objekten (Named Entity Recognition – NER) zum Einsatz. Diese wurde gemäß der UIMA-Architektur modular gestaltet. Als Basis für die NER kommt ein Wörterbuch zum Einsatz, das direkt aus der Anwendungs-Ontologie heraus generiert wird. Diese enthält bereits übliche Bezeichnungen in mehreren Sprachen sowie Synonyme für jedes modellierte Konzept. Hier zeigt sich der Mehrwert der semantischen Modellierung: Das gleiche Modell kann sowohl für die Verknüpfung von Informationen, für die Disambiguierung von Nutzeranfragen sowie für die Informationsextraktion eingesetzt werden.

Für die Anbindung strukturierter Quellen werden zwei Möglichkeiten genutzt. Zunächst bietet OntoBroker ein direktes Mapping von SQL-Datenbanken an. Für andere strukturierte Quellen wird XML in Verbindung mit XSL-Transformationen eingesetzt. Daten werden in XML exportiert und dann in RDF/XML transformiert. Dabei werden gleichzeitig XML-Elemente auf entsprechende Konzepte in der Anwendungs-Ontologie abgebildet. Die Abbildung muss bislang noch für jede anzubindende Quelle manuell erstellt werden. Es ist jedoch geplant, generische Mapping-Mechanismen einzusetzen.

4 Zusammenfassung und Ausblick

In diesem kurzen Beitrag haben wir die Vision des föderierten Produktinformationssystems sowie die beiden wichtigsten Technologien Semantic Computing und Informationsextraktion am Beispiel des Projektes Aletheia vorgestellt. Obwohl beide Technologien im Semantic Web bzw. beim Unstructured Information Management in den letzten Jahren immer mehr zum Einsatz kommen, werden ihre Stärken bislang selten kombiniert. Aufgrund der positiven Erfahrungen im Rahmen des Projektes Aletheia sind wir überzeugt, dass die Zusammenführung beider Technologien ein großes Potential bietet.

Weitere Forschungsarbeit wird benötigt, um die Genauigkeit der Ontologie-basierten Informationsextraktion zu verbessern. Eine Kombination mit Trainings-basierten Ansätzen scheint hier der richtige Weg zu sein. Weitere Arbeiten sind im Bereich der dynamischen Anbindung von Quellen sowie dem Änderungs-Management geplant. Umfangreiche Nutzerstudien in der letzten Phase des Projektes sollen die Vorteile des hier gezeigten Ansatzes untermauern.

Literaturverzeichnis

[1] Aperture - a java framework for getting data and metadata, http://aperture.sourceforge.net/, 2010.

[2] Aduna, openrdf.org: Home, http://www.openrdf.org/, 2010.

[3] Ontoprise GmbH, ontoprise: OntoBroker, http://www.ontoprise.de/deutsch/start/produkte/ontobroker/, 2010.

[4] Ontoprise GmbH, ontoprise: OntoStudio, http://www.ontoprise.de/deutsch/start/produkte/ontostudio/, 2010.

[5] Seth Grimes, Unstructured data and the 80 percent rule - investigating the 80%, Technical report, Clarabridge, 2008.

[6] Aletheia Konsortium, Aletheia - Semantische Föderation umfassender Produktinformationen, http://aletheia-projekt.de/, 2010.

[7] Sandro Reichert, A Secure Data Repository for Semantic Federation of Product Information, 11th International Conference on Information Integration and Web-based Applications & Services (iiWAS2009), Kuala Lumpur, Malaysia, 2009.

[8] Matthias Wauer, Daniel Schuster, Johannes Meinecke, Aletheia - An Architecture for Semantic Federation of Product Information from Structured and Unstructured Sources, 12th International Conference on Information Integration and Web-based Applications & Services (iiWAS2010), Paris, France, 2010.

[9] Andrew White, Magic quadrant for master data management of product data, Technical report, Gartner Research, 2009.

Danksagung: Die hier beschriebenen Arbeiten wurden teilweise mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IA08001 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.