Schreier G, Hayn D, Hörbst A, Ammenwerth E, editors. Proceedings of the eHealth2012. 2012 Mai 10-11; Vienna, Austria. OCG; 2012.

EIN ONTOLOGIEBASIERTES SYSTEM ZUM EXTRAHIEREN, TRANSFORMIEREN UND LADEN VON DATEN IN KRANKENANSTALTEN

Arthofer K[1], Girardi D[2], Giretzlehner M[2]

Kurzfassung

Adäquate Datenqualität ist eine Voraussetzung für valide Auswertungen. Leider stellt gerade semantische Datenqualität in Krankenanstalten ein großes Problem dar. Ein von den Autoren in einem aktuellen Forschungsprojekt entwickeltes semantisches ETL-System namens OBIK hat in einem Pilotprojekt bereits gezeigt, dass es die Datenqualität in Krankenanstalten im Rahmen von Auswertungen verbessern kann. Der Artikel differenziert das Problem der Datenqualität in Krankenanstalten und erklärt die Architektur und die Anwendung von OBIK.

Abstract

Valid data analysis strongly depends on the quality of the underlying data. Unfortunately, hospitals are facing massive drawbacks especially when it comes to semantic data quality. We developed a semantic ETL system called OBIK, which could already proof its ability to improve semantic data quality for data analysis. We point out the problem of data quality in hospitals and describe the OBIK system in detail. Furthermore, we provide our conclusions and future outlook of the project.

Keywords – ETL, Datenqualität, Metadatenmodell

1. Einleitung

Extrahieren, Transformieren und Laden von Daten (ETL) ist in Krankenanstalten insbesondere im Zusammenhang von Data Warehouses (DWH) etabliert. Der Hauptvorteil liegt im zentralen Langzeitspeichern aggregierter, integrierter und damit konsolidierter Daten. Es ist altbekannt, dass das ETL aus technischer Hinsicht, die technische Installation eines DWH uä. nur die Spitze des Eisbergs sind. Der unter der „Wasseroberfläche" befindliche Teil des oft zitierten DWH-Eisbergs besteht aus Erfolgsfaktoren wie in der Organisation anerkannter Unternehmensziele sowie damit verbundener Auswertungsziele, Konsolidierung betriebswirtschaftlicher Methoden und anderer abstrakter Sachverhalte. Ein im Vergleich konkreter Erfolgsfaktor unter der „Wasseroberfläche" ist die semantische Datenqualität, also nicht Datenvalidität bzgl. Dateiformat, Datenformat oder anderen technischen Aspekten sondern Validität bzgl. der Bedeutung der Daten [1].

Obwohl das Problem Datenqualität gravierend und auch altbekannt ist, ist es nach wie vor ein aktuelles bzw. aufgeschobenes Problem mit schwerwiegenden Folgen für die Brauchbarkeit von Auswertungen („garbage in, garbage out"). Eine aktuelle, internationale Studie zum Stammdaten

[1] Studiengang Prozessmanagement Gesundheit – FH Oberösterreich Campus Steyr
[2] RISC Software GmbH – Research Unit Medical Informatics

Management von pwc Deutschland reiht Dienstleistungsbetriebe und damit auch Krankenanstalten im Branchenvergleich bzgl. Datenqualität an die letzte Stelle. Diese Studie sieht auch Datenqualität untrennbar mit Stammdaten Management verbunden und sieht sie bildhaft als „zwei Seiten derselben Medaille“ [2]. Der Zusammenhang von Datenqualität und Stammdaten Management wird z.B. auch am Institut für Wirtschaftsinformatik der Universität St. Gallen entsprechend gesehen [3, 4]. Auch ein im Rahmen des von der Deutschen Gesellschaft für Informations- und Datenqualität (DGIQ e.V.) bei der CeBIT 2007 gestalteten Thementages erschienener Artikel ortet mangelhafte Datenqualität in Krankenhäusern und weist auf zahlreiche dadurch ungenutzte Potentiale (Optimierung des Controlling, Prozessen, Business Intelligence-Systeme etc.) hin [5]. Schließlich haben die Autoren dieses Artikels in ihren zahlreichen Projekten mit Krankenanstalten ebenfalls entsprechende Erfahrungen gemacht.

Erschwerend kommt beim Problem der Datenqualität hinzu, dass es dynamisch ist. Einerseits gewinnt man bei der Auswertung von Daten nicht nur Erkenntnisse über den Gegenstand der jeweiligen Daten (Deckungsbeiträge, Komplikationsraten medizinischer Behandlungen etc.) sondern auch über die Beschaffenheit der Daten selbst. So beschließt man z.B. nach einer Auswertung, zusätzliche Daten elektronisch zu erfassen, um differenziertere Auswertungen vornehmen zu können. Andererseits unterliegt jede Organisation und damit auch ihre Datenstrukturen einer permanenten Veränderung. Das heißt, man muss beim ETL von Daten einer Domäne über die Zeit auch die Veränderung von Datenstrukturen bewerkstelligen. Man kann das ETL und Auswerten von Daten nicht als einmaligen Aufwand betrachten sondern muss es als Zyklus und kontinuierlichen Lernprozess verstehen. Schließlich macht es auch oft Sinn, im Volumen geringe jedoch in deren Relevanz sehr bedeutende aber leider unstrukturierte oder überhaupt analoge Datenbestände im Rahmen des ETL ergänzend manuell zu erfassen. Jedenfalls werden Auswertungen erst über Zeit richtig aussagekräftig [6].

Angesichts der hier geschilderten Problematik entwickeln die Autoren im Rahmen eines aktuellen Forschungsprojektes eine Ontologie-basierte Benchmarking-Infrastruktur für Krankenanstalten (OBIK) welche bereits für einen Referenz-Krankenhausträger eingesetzt wird und bereits erste Auswertungen geliefert hat.

2. Generische Datenmodelle

2. 1. Konventionelle datenbasierte Softwaresysteme

Moderne datenbasierte Softwaresysteme sind in mehreren logischen Schichten organisiert [7]. Diese Auftrennung des Systems in meist drei Schichten (*three tier architecture* bestehend aus Benutzeroberfläche, Fachkonzept, Datenhaltung) ermöglicht nicht nur die Verteilung der Applikation in Client-Server-Architekturen; sie reduziert auch die Abhängigkeiten zwischen den einzelnen Schichten. So kann beispielsweise die Benutzeroberfläche ausgetauscht oder erweitert werden, ohne dass die darunterliegenden Schichten davon betroffen sind.

Obwohl Mehrschichtarchitekturen dazu beitragen, technische Abhängigkeiten zu minimieren, sind sie trotzdem nicht in der Lage die Abhängigkeit der gesamten Applikation von der Datenstruktur der realen Problemstellung (*real world problem*) zu lösen. Das gesamte System hängt sehr stark von der Struktur der Domäne ab, für die das System im Einsatz ist. Werden Daten für eine medizinische Studie gesammelt, so haben diese eine komplett andere Struktur, als zum Beispiel bei der Erfassung von meteorologischen Daten. Datenerfassungssysteme für die beiden Anwendungsfälle wären in ihrem semantischen Aufbau grundverschieden und könnten in anderen Domänen nicht eingesetzt werden. Besonders im medizinischen Bereich mit seinen Fächern ist eine weit differenzierte Spezifikation der Datenstrukturen notwendig. Sogar innerhalb eines Faches kann

es auf Grund verschiedener Forschungsfragen oder unterschiedlicher Institutionen zu signifikant unterschiedlichen Anforderungen an die Datenstruktur und die semantische Überprüfung kommen. All diese verschiedenen Fachdomänen in einem globalen System zu erfassen wäre daher nicht zielführend und würde den Einsatz von mehreren domänenspezifischen Lösungen notwendig machen. Dies würde die Entwicklung mehrere Systeme erfordern, was sehr kostspielig und zeitaufwendig wäre. Aber nicht nur die Einrichtung solcher Systeme wäre sehr aufwendig, auch deren Anpassung an sich ändernde Anforderungen. Um also eine technische Infrastruktur zu schaffen, mit der man in der Lage ist, Daten verschiedenster Struktur für unterschiedlichste Anwendungsbereiche zu erfassen und das sich schnell an geänderte Anforderungen anpassen lässt, muss die semantische Abhängigkeit der Datenerfassung von der Anwendungsdomäne umgangen werden.

Um diese Abhängigkeit aufzulösen, müssen sowohl die Datenhaltungs- und die Fachkonzeptschicht in der Lage sein, generische Konzepte höherer Abstraktionsebenen zu verarbeiten. Herkömmliche Datenmodelle, wie sie in der Datenhaltungsschicht von Dreischichtarchitekturen verwendet werden, beschreiben die Anwendungsdomäne direkt. Metadatenmodelle befinden sich eine Abstraktionsstufe über diesen direkten Datenmodellen und sind Datenmodelle, die in der Lage sind, andere Datenmodelle zu beschreiben, zu speichern und auch deren Daten zu speichern.

Durch die Implementierung eines Metadatenmodells in der Datenhaltungsschicht ist die Anwendung nun in der Lage, Daten beliebiger Strukturen und Informationen über eben diese Strukturen zu speichern. Das Metadatenmodell ist domänenunabhängig. Die Anwendungsdomäne bestimmt nun nicht mehr die Struktur der Datenhaltungsschicht sondern nur mehr deren Inhalt. Man spricht hier von einer *Instanziierung* des Metadatenmodells durch ein domänenabhängiges Datenmodell. Die Abhängigkeit der Anwendung zur Anwendungsdomäne ist somit aufgelöst. Um eine Fachkonzeptebene zu erhalten, die ebenfalls in der Lage ist, unabhängig von dem instanziierten Datenmodell zu arbeiten, darf die Logik dieser Ebene nicht mehr direkt im Sourcecode implementiert, sondern muss ebenfalls im Metadatenmodell hinterlegt sein. Metadatenmodelle, die sowohl Strukturinformationen, als auch Bewegungsdaten gemeinsam mit Logikinformation beinhalten werden auch als Ontologie bezeichnet.

2. 2. Ontologie-basierte Datenhaltung

Für den Begriff Ontologie gibt es in der Informatik zahlreiche Definitionen. Chandrasekaran et al [8] definieren Ontologie wie folgt: ”*Ontologies are content theories about the sorts of objects, properties of objects, and relations between objects that are possible in a specified domain of knowledge.*”. Gruber [10] bleibt noch vager und bezeichnet Ontologien lediglich als: “*An ontology is a specification of a conceptualization*.”

Ontologie-basierte oder Metamodell-basierte Systeme (beide Begriffe werden hier synonym verwendet) zeichnen sich dadurch aus, dass sie in der Lage sind, Daten beliebiger Struktur zu erfassen. Die zum Datenmodell gehörige Logik wird ebenfalls in der Ontologie definiert. Somit wird das für die Datenerfassung notwendige und formalisierbare Domainwissen verwendet um die generische Applikation zu instanziieren und somit für die Verwendung vorzubereiten. Die Datenhaltungsschicht beinhaltet das generische Metadatenmodell, welches das aktuelle Instanzmodell speichert. Die Benutzeroberfläche(n) kann (können) aus dem im Metamodell gespeicherter Metainformationen automatisch (zur Laufzeit) erzeugt werden, wobei zahlreiche optionale Konfigurationsmöglichkeiten die Individualisierung der Formulare erlauben. Dies ist nun ein weiterer Vorteil dieser Systeme: Ändert sich das Instanzdatenmodell so müssen bei einer konventionellen Software zahlreiche (oft langwierige und damit kostspielige) Änderungen vorgenommen werden. Bei einer Metamodell-basierten Software propagieren sich die Änderungen

automatisch fort. Die Benutzeroberfläche wird abhängig vom Instanzmodell zur Laufzeit erzeugt. Ändert sich dieses, so wird beim nächsten Aufruf automatisch eine neue Oberfläche aufbauend auf den aktuellen Strukturinformationen erzeugt.

2. 3. Ontologie-basierte Plausibilitätsprüfungen

Ontologien werden nicht nur verwendet um Datenmodelle zu definieren. Sie beinhalten die teils komplexen Regeln für die semantische Überprüfung der Daten bezüglich ihrer Plausibilität. In einem Ontologie-basierten Datenerfassungssystem ist dies notwendig um die Domänenunabhängigkeit, die in der Datenhaltungsschicht durch das Metadatenmodell gegeben ist, auch in der Fachkonzeptschicht zu erreichen. Obwohl es aus Sicht des Entwicklers umständlich erscheinen mag, die Logik über ein Regelwerk zu definieren anstatt direkt zu implementieren, ergeben sich neben der Domänenunabhängigkeit eine Reihe weiterer Vorteile:

- Um das Regelwerk zu verwalten ist nun kein Zugriff auf den Sourcecode des Systems notwendig. Dies ermöglicht es nicht nur das Regelwerk jederzeit zu ändern und zu erweitern, sondern erlaubt auch Personen, die über keinerlei Programmierkenntnisse (dafür aber über domänenspezifisches Fachwissen) verfügen am Regelwerk zu arbeiten.
- Durch die explizite Definition der Regeln können diese übersichtlich dargestellt und verwaltet werden. Grafische Frameworks können verwendet werden um die Abhängigkeiten im System zu visualisieren.
- Regeln die in strukturierter Weise gespeichert werden, können auch auf Ihre Gültigkeit hin bzw. auf Widersprüche automatisch untersucht werden.

Die Fachkonzeptebene besteht nun lediglich aus einem Interpreter, der die Regeln aus der Ontologie ausliest und diese auf die Daten anwendet. Dieser generische Regelinterpreter ist das Pendant zum generischen Metadatenmodell und hebt die Fachkonzeptebene auf ein höheres Abstraktionsniveau.

3. OBIK

Um den in Abschnitt 1 beschriebenen Herausforderungen gerecht zu werden, wurde eine Ontologie-basierte Datenerfassungs- und Plausibilisierungsinfrastruktur entwickelt. Dadurch kann die erforderte Flexibilität erreicht werden, während gleichzeitig ein komplexes ebenfalls generisches Regelwerk die Plausibilität der Daten überprüft.

3. 1. Systemaufbau

Das Softwaresystem besteht aus fünf Hauptkomponenten:

1. Als Datenhaltungsschicht dient ein relationales Datenbanksystem mit einem generischen Metadatenmodell. Für technische Details zum Metadatenmodell wird auf folgendes Paper verwiesen: Girardi et al [9].
2. Darauf aufbauend wurde ein Admin-Client entwickelt mit dem die Autoren in der Lage sind, die Anwendungsontologie mit dazugehörigem Regelwerk zu modellieren und zu warten, sowie die im System befindlichen Bewegungsdaten zu bearbeiten.
3. Um die Überführung von elektronisch gespeicherten Daten weitest möglich zu automatisieren, wurde ein ETL-Framework und ein Ontologie-Modul entworfen, welches in der Lage ist, Quelldaten auf die aktuell instanziierte Zieldatenstruktur zu überführen.
4. Um es speziell geschulten Erfassungskräften zu ermöglichen elektronisch importierte Daten mit handschriftlich erfassten Daten zu ergänzen, wurde eine Webschnittstelle entwickelt,

welche zur Laufzeit basierend auf den Metainformationen aus der Ontologie aufgebaut wird. Dort können die Erfassungskräfte die schon vorhandenen Daten im Webbrowser betrachten, editieren und ergänzen.

5. Plausibilisierung: Nachdem der Studienautor die Datenstrukturen definiert hat und die Daten über verschiedene Wege ins System importiert wurden, können sie auf ihre Plausibilität hin überprüft werden. Hierbei stehen mehrere Arten von Plausibilitätsüberprüfungen zur Verfügung, die mittels logischer Operatoren (UND und ODER) zu beliebig komplexen logischen Ausdrücken kombiniert werden können. All diese Regel können über eine intuitiv gestaltete Benutzeroberfläche erstellt und verwaltet werden. Dafür sind keinerlei Programmier- oder Logikkenntnisse notwendig. Die Abhängigkeiten der einzelnen Attribute, die der Benutzer definiert, werden vom System in einem Graph übersichtlich dargestellt.

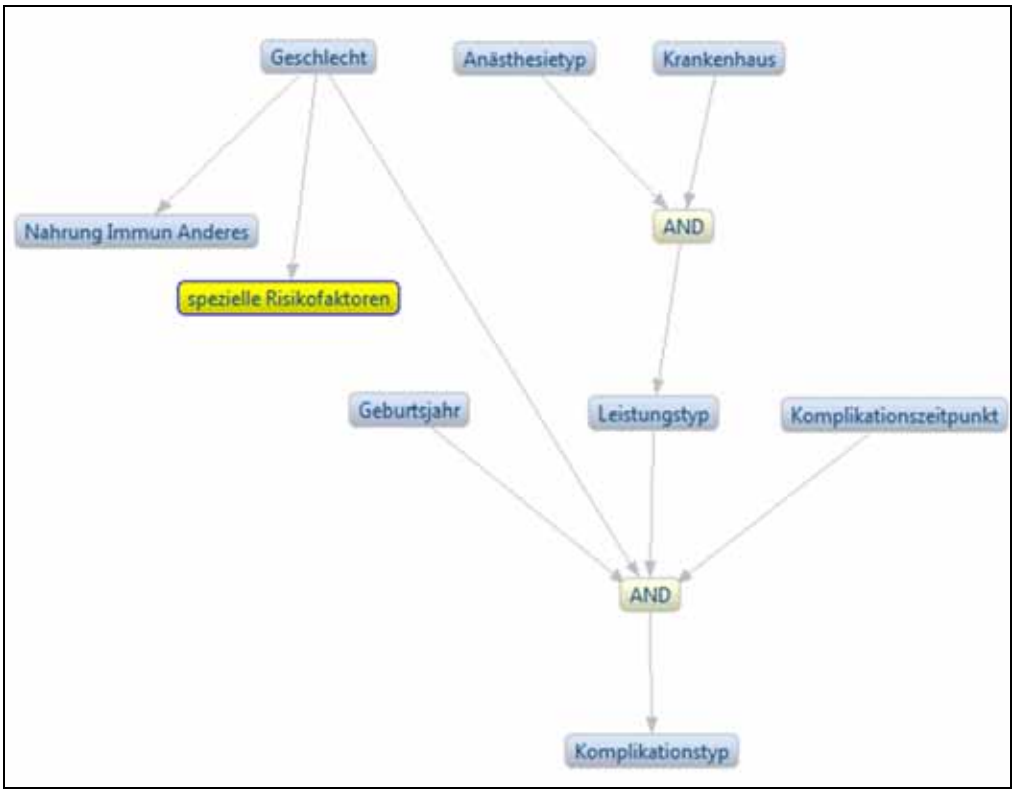

Abbildung 1. Darstellung der Abhängigkeiten

Abbildung 1 zeigt einen solchen Graph, in dem die Abhängigkeiten, die im Rahmen einer Plausibilitätsprüfung überprüft werden, zu sehen sind. Hier sieht man, dass beispielsweise der Leistungstyp nur plausibel ist, wenn der Anästhesietyp und das dazugehörigen Krankenhaus entsprechende Werte aufweisen. Die Plausibilität einer Komplikation hängt wiederum von einer Kombination von mehreren Faktoren ab wie Geschlecht des Patienten, sein Geburtsjahr, die erbrachte medizinische Leistung und der Zeitpunkt, wann diese Komplikation aufgetreten ist. Eine Verletzung des Eierstockes, bei einem männlichen Patienten nach einer Nasenkorrektur wäre ein Beispiel für die Verletzung gleich mehrerer solcher Regeln.

Als Ergebnis der Überprüfung liefert das System eine Liste der Datensätze, die der Überprüfung nicht standgehalten haben mit detaillierter Information welcher Wert, welche Regel auf welche Art und Weise verletzt. Zusätzlich zu dieser Liste werden die Fehlermeldungen in der Datenbank,

assoziiert mit den jeweiligen Datensätzen, abgelegt. So können die Erfassungskräfte beim Öffnen des jeweiligen Datensatzes die dazugehörigen Fehler auf einen Blick erkennen.

4. Schlussfolgerungen

Mit OBIK konnte ein ETL-System entwickelt werden, das auf Basis einer entsprechenden domänenspezifischen Ontologie sowohl eine Datenstruktur für Auswertungen als auch eine Web-Oberfläche zur manuellen Erfassung ergänzender Daten generiert und syntaktische und insbesondere semantische Plausibilitätsprüfungen auswerten kann. Der entwickelte Admin-Client ermöglicht es auch Benutzern ohne Datenbank- und Programmierkenntnissen die domänenspezifische Ontologie zu erstellen und somit das Datenerfassungssystem inkl. Plausibilisierung innerhalb kürzester Zeit für den Einsatz vorzubereiten. Des Weiteren lässt sich mit Hilfe von OBIK die domänenspezifische Ontologie mit geringem Aufwand an veränderte Bedingungen anpassen. Damit kann OBIK die Grundlage für valide Auswertungen in Krankenanstalten gewährleisten. Dadurch wird eine Grundlage für alle auf Datenauswertungen basierende Aktivitäten wie z.B. Qualitätsmanagement und Controlling gebildet und kann auch in der Abrechnung nach Leistungsorientierter Krankenanstaltenfinanzierung (LKF) Vorteile bringen. Schließlich ermöglicht OBIK durch seine Agilität eine kontinuierliche Verbesserung der den Auswertungen zugrundeliegenden Daten, der Auswertungen selbst und ggf. auch der Datenorganisation (vgl. Stammdaten Management) einer Organisation.

Bei dem in der Einleitung bereits erwähnten Referenz-Krankenhausträger wurden mit OBIK bis dato rund 2000 Benchmarking-Fälle der Behandlungsfallklassen „Appendix“, „Hernie“, „Struma“ und „Galle“ aus dem 4. Quartal 2010 und 1. Quartal 2011 aufbereitet. Mit diesen Fällen wurden 30 % des Behandlungsvolumens der chirurgischen Abteilungen des Referenz-Krankenhausträgers durch finanzielle und medizinische Ergebniskennzahlen transparent. Voraussetzung für das Erheben dieser Benchmarking-Fälle war natürlich einerseits das Konsolidieren der relevanten Stammdaten (insbes. Hausleistungskataloge) bzw. genau genommen die für die adressierten Behandlungsfallklassen relevanten Subkataloge. Andererseits mussten auch die manuell ergänzten (aktuell 25) Datenelemente (z.B. Risikofaktoren) initial eindeutig und konsistent definiert werden. Etwa 1000 dieser Benchmarking-Fälle wurden doppelt erfasst (so gesehen wurden also insgesamt 3000 in der Basis elektronisch eingelesener Fälle mit der Web-Oberfläche von OBIK ergänzt). sind In Ergänzung zur Doppelerfassung sind in OBIK zurzeit insgesamt etwa 200 Plausibilitätsregeln unterschiedlicher Komplexität abgebildet. Die Prüfung dieser Plausibilitätsprüfungen legte vielfältige Datenprobleme wie z.B. falsche Behandlungszeiten (Leistungsdatum), fehlende Daten, vereinzelt mangelhaft codierte LKF-Diagnosen und -Leistungen usw. offen. Parallel dazu ergab die Datenextraktion aus Papierakten bzw. der Abgleich der elektronischen Daten fehlende Befunde und inkonsistent dokumentierte und damit schwer auffindbar erfasste Daten. Da diese Erkenntnisse zur Verbesserung sowohl der elektronischen als auch papierenen Dokumentation verwendet werden können, dient OBIK dem Referenz-Krankenhausträger nicht nur zur Aufbereitung seiner Daten sondern auch allgemein zum Management seiner Daten. Aufgrund der in der Einleitung erwähnten Studien bzw. der eigenen Erfahrung zu Datenqualität in Krankenanstalten ist davon auszugehen, dass OBIK in jedem Krankenhaus entsprechendes Potential bietet.

5. Literatur

[1] Kurz A. Data Warehousing. Enabling Technology, mitp-Verlag 1999, pp. 113-115.

[2] Messerschmidt M, Stüben J. Verborgene Schätze. Eine internationale Studie zum Master-Data-Management, pwc Deutschland, 2011, pp. 76-77.

[3] Boris O. Data Governance, In: Wirtschaftsinformatik 4/2011, Gabler, pp. 235-238.

[4] Boris O. Stammdatenmanagement: Datenqualität für Geschäftsprozesse. In: Praxis der Wirtschafsinformatik, HMD 279, dpunkt.Verlag, 2011, pp. 5-16.

[5] Hüfner J. Datenqualität im Krankenhaus. Kostenvorteile durch ausgereifte Konzepte, http://www.tiq-solutions.de/download/attachments/425996/Datenqualitaet-im-Krankenhaus_Jan-Huefner_08-2007.pdf (24.01.2012)

[6] Bell S, Orzen M. Lean IT. Enabling and Sustaining Your Lean Transformation, Productivity Press, New York 2011, pp. 133-135.

[7] Balzert H. Java: objektorientiert programmieren. Vom objektorientierten Analysemodell bis zum objektorientierten Programm. 2nd ed. Herdecke, Witten: W3L-Verl. 2010

[8] Chandrasekaran B, Josephson J R, Benjamins V R. What are Ontologies, and why do we need them? IEEE Intelligent Systems 14, 1 (Jan./Feb. 1999), 20-26

[9] Girardi D, Dirnberger J, Giretzlehner M. Meta-model based knowledge discovery. 2011 International Conference on Data and Knowledge Engineering (ICDKE). Milan. 2011

[10] Gruber T. A translation approach to portable ontologies. Knowledge Acquisition, 5(2):199-220. 1992

Corresponding Author
Klaus Arthofer
Studiengang Prozessmanagement Gesundheit
FH Oberösterreich Campus Steyr
klaus.arthofer@fh-steyr.at