Evolution von ontologiebasierten Mappings in den Lebenswissenschaften

Zusammenfassung der Dissertation

Anika Groß, Abteilung Datenbanken am Institut für Informatik, Universität Leipzig

Im Bereich der *Lebenswissenschaften* kommen *Ontologien* und andere strukturierte Vokabulare zum Einsatz, um eine einheitliche Erfassung von Wissen sowie einen formalen Austausch zwischen verschiedenen Applikation zu erleichtern. Ontologien finden Anwendung in verschiedenen Domänen wie der Molekularbiologie oder Chemie und dienen zumeist der *Annotation* realer Objekte wie z.B. Gene oder Literaturquellen. Unterschiedliche Ontologien enthalten jedoch teilweise überlappendes Wissen, so dass die Bestimmung einer Abbildung (*Ontologiemapping*) zwischen ihnen notwendig ist. Oft ist eine manuelle Mappingerstellung zwischen großen Ontologien kaum möglich, weshalb typischerweise automatische Verfahren zu deren Abgleich (*Matching*) eingesetzt werden. Aufgrund neuer Forschungserkenntnisse und Nutzeranforderungen entwickeln sich die Ontologien kontinuierlich weiter. Die Evolution der Ontologien hat wiederum Auswirkungen auf abhängige Daten wie beispielsweise Annotations- und Ontologiemappings, welche entsprechend aktualisiert werden müssen. Im Rahmen dieser Arbeit werden neue Methoden und Algorithmen zum Umgang mit der *Evolution ontologiebasierter Mappings* entwickelt. Dabei wird die generische Infrastruktur *GOMMA* zur Verwaltung und Analyse der Evolution von Ontologien und Mappings genutzt und erweitert.

Zunächst wurde eine vergleichende *Analyse der Evolution* von Ontologiemappings für drei Subdomänen der Lebenswissenschaften durchgeführt. Insgesamt zeigt sich ein deutlicher Einfluss von Ontologieänderungen auf Ontologiemappings. Dementsprechend können bestehende Mappings infolge der Weiterentwicklung von Ontologien ungültig werden, so dass sie auf aktuelle Ontologieversionen migriert werden müssen. Dabei sollte eine aufwendige Neubestimmung der Mappings vermieden werden. In dieser Arbeit werden zwei generische Algorithmen zur (semi-) automatischen *Adaptierung* von Ontologiemappings eingeführt. Ein Ansatz basiert auf der Komposition von Ontologiemappings, wohingegen der andere Ansatz eine individuelle Behandlung von Ontologieänderungen zur Adaptierung der Mappings erlaubt. Beide Verfahren ermöglichen die Wiederverwendung unbeeinflusster, bereits bestätigter Mappingteile und adaptieren nur die von Änderungen betroffenen Bereiche der Mappings. Eine Evaluierung für sehr große, biomedizinische Ontologien und Mappings zeigt, dass beide Verfahren qualitativ hochwertige Ergebnisse produzieren.

Ähnlich zu Ontologiemappings werden auch ontologiebasierte *Annotationsmappings* durch Ontologieänderungen beeinflusst. Die Arbeit stellt einen generischen Ansatz zur Bewertung der Qualität von Annotationsmappings auf Basis ihrer Evolution vor. Verschiedene Qualitätsmaße erlauben die Identifikation glaubwürdiger Annotationen beispielsweise anhand ihrer *Stabilität* oder Herkunftsinformationen. Eine umfassende Analyse großer Annotationsdatenquellen zeigt zahlreiche Instabilitäten z.B. aufgrund temporärer Annotationslöschungen. Dementsprechend stellt sich die Frage, inwieweit die Datenevolution zu einer Veränderung von abhängigen Analyseergebnissen führen kann. Dazu werden die Auswirkungen der Ontologie- und Annotationsevolution auf sogenannte funktionale Analysen großer biologischer Datensätze untersucht. Eine Evaluierung anhand verschiedener Stabilitätsmaße erlaubt die Bewertung der Änderungsintensität der Ergebnisse und gibt Aufschluss, inwieweit Nutzer mit einer signifikanten Veränderung ihrer Ergebnisse rechnen müssen.

Darüber hinaus wird GOMMA um *effiziente Verfahren* für das *Matching sehr großer Ontologien* erweitert. Diese werden u.a. für den Abgleich neuer Konzepte während der Adaptierung von Ontologiemappings benötigt. Viele der existierenden Match-Systeme skalieren nicht für das Matching besonders großer Ontologien wie sie im Bereich der Lebenswissenschaften auftreten. Ein effizienter, *kompositionsbasierter Ansatz* gleicht Ontologien indirekt ab, indem existierende Mappings zu

Mediatorontologien wiederverwendet und miteinander kombiniert werden. Zudem werden generelle Strategien für das *parallele Ontologie-Matching* unter Verwendung mehrerer Rechenknoten vorgestellt. Eine größenbasierte Partitionierung der Eingabeontologien verspricht eine gute Lastbalancierung und Skalierbarkeit, da kleinere Match-Teilaufgaben parallel verarbeitet werden können. GOMMA erreichte zur Evaluierung im Rahmen der *Ontology Alignment Evaluation Initiative* (OAEI) 2012 sehr gute Ergebnisse bezüglich der Effektivität und Effizienz des Matchings, insbesondere für Ontologien aus dem Bereich der Lebenswissenschaften.