

Evolution von Ontologien in den Lebenswissenschaften

Zusammenfassung der Dissertation

Michael Hartung

*Interdisziplinäres Zentrum für Bioinformatik und Abteilung Datenbanken am Institut für Informatik
Universität Leipzig*

In den *Lebenswissenschaften* haben sich *Ontologien* in den letzten Jahren auf breiter Front durchgesetzt und sind in vielen Anwendungs- und Analyseszenarien kaum mehr wegzudenken. So etablierten sich nach und nach immer mehr domänenspezifische Ontologien, z.B. Anatomie-Ontologien oder Ontologien zur Beschreibung der Funktionen von Genen oder Proteinen. Da das Wissen in den Lebenswissenschaften sich rapide ändert und weiterentwickelt, müssen die entsprechenden Ontologien ständig angepasst und verändert werden, um einen möglichst aktuellen Wissensstand zu repräsentieren. Nutzer von Ontologien müssen mit dieser *Evolution* umgehen können, d.h. um "Up-to-Date" zu sein, sollten die aktuellsten Versionen einer Ontologie verwendet werden. Dies ist häufig nur schwer umsetzbar, da die Evolution weitreichende Einflüsse auf existierende Datenbestände, Analyseergebnisse oder Anwendungen haben kann. Innerhalb dieser Dissertation stehen Werkzeuge und Algorithmen zum Umgang mit sich ständig ändernden Ontologien im Bereich der Lebenswissenschaften im Mittelpunkt.

Zunächst wird ein generelles Framework für *quantitative Evolutionsanalysen* eingeführt. Das Framework wird für eine umfassende Analyse der Evolution zahlreicher Ontologien der Lebenswissenschaften verwendet. Die Analysen zeigen, dass alle untersuchten Ontologien stetig verändert (angepasst) werden und ein signifikantes Wachstum aufweisen. Auch für auf Ontologien basierte Mappings, d.h. Verknüpfungen zwischen Datenquellen und Ontologien (Annotation-Mapping) sowie zwischen Ontologien selbst (Ontologie-Mapping), liegen starke und häufige Veränderungen vor. Es besteht somit ein Bedarf, die Evolution von Ontologien in den Lebenswissenschaften und deren Konsequenzen zu unterstützen, d.h. Nutzern von sich ständig ändernden Ontologien angemessene Algorithmen/Werkzeuge bereitzustellen. Die Erkenntnisse aus den durchgeführten Analysen bilden die Basis für die nachfolgenden Arbeiten.

Eine immer wiederkehrende Aufgabe im Rahmen der Ontologieevolution besteht in der Bestimmung von Änderungen zwischen zwei Versionen einer Ontologie, d.h. worin besteht der Unterschied und wie hat sich die neuere Version aus der alten Version heraus entwickelt. Das Ergebnis, d.h. der *Diff* (die Differenz) zwischen den beiden Ontologieversionen, bildet die Basis für weitere Aufgaben wie beispielsweise die Anpassung abhängiger Daten. Innerhalb der Arbeit wird ein neuartiger auf Regeln basierender Algorithmus vorgestellt, welcher den Diff zwischen zwei Ontologieversionen bestimmt. Es werden sowohl einfache wie auch komplexe Änderungen erkannt, was eine kompakte, intuitive und verständliche Diff-Repräsentation garantiert. Es wird theoretisch wie praktisch gezeigt, dass ein vollständiger Diff bestimmt wird, was eine korrekte Migration von Ontologieversionen ermöglicht.

Ein weiterer Schwerpunkt der Arbeit liegt in der Bestimmung änderungsintensiver bzw. stabiler Regionen in einer Ontologie. Dazu wird die Notation von *Ontologieregionen* und zugehörige Metriken zur Beurteilung ihrer Änderungsintensität (Stabilität) eingeführt. Ein neuartiger automatisierter Algorithmus erlaubt die Bestimmung (in)stabiler Ontologieregionen auf Basis veröffentlichter Ontologieversionen in einem vorgegebenen Zeitraum. Durch erkannte Änderungen zwischen Ontologieversionen und mit Hilfe der Ontologiestruktur werden änderungsintensive bzw. stabile Ontologieregionen erkannt. Die Evaluierung anhand großer Ontologien der Lebenswissenschaften zeigt, dass der Algorithmus in der Lage ist (in)stabile Ontologieregionen automatisiert zu bestimmen.

Abschließend wird das webbasierte System *OnEX* und dessen *Versionierungsansatz* präsentiert. *OnEX* ermöglicht einen benutzerfreundlichen und interaktiven Zugang zu Informationen über die Evolution und Änderungen in Ontologien der Lebenswissenschaften. Nutzer können Ontologien aus ihrem Interessengebiet bzgl. Evolution untersuchen, indem sie beispielsweise Änderungen an einer Ontologieversion einsehen, welche in einer Analyse oder Anwendung genutzt werden soll. Der *OnEX* zugrunde liegende Versionierungsansatz ermöglicht eine skalierbare und speichereffiziente Versionierung großer Ontologien durch die Nutzung von Zeitstempeln. Mit Hilfe des Ansatzes konnten 16 Ontologien mit ca. 700 Versionen seit 2002 versioniert und Nutzern über *OnEX* für Evolutionsanalysen zugänglich gemacht werden.