Management von Ontologien in den Lebenswissenschaften

Michael Hartung

Interdisziplinäres Zentrum für Bioinformatik
Universität Leipzig
hartung@izbi.uni-leipzig.de

Zusammenfassung

Die Bedeutung bzw. der praktische Nutzen von Ontologien zeigt sich insbesondere in den Lebenswissenschaften. Durch die gestiegene Akzeptanz und Anwendung von Ontologien stellt deren Management mehr und mehr ein wichtiges Problem dar. Aus diesen Grund werden die ständige Weiterentwicklung (Evolution) von Ontologien und die starke Heterogenität bezüglich ihrer Formate in diesem Beitrag näher betrachtet. Im speziellen versucht ein erster Ansatz verschiedene Ontologien über eine Middleware in Gridumgebungen zu integrieren, um Gridnutzern bzw. Anwendungen im Grid einen einheitlichen und einfachen Zugang zu Ontologieinformationen zu ermöglichen. Eine weitere Untersuchung beschäftigt sich primär mit der Evolution von Ontologien. Auf Basis eines Frameworks wurde eine quantitative Analyse der Evolution von 16 aktuell verfügbaren, biomedizinischen Ontologien durchgeführt.

1 Einführung

Ontologien erleben insbesondere in den Lebenswissenschaften eine stetig steigende Anwendung und Bedeutung. Sie beinhalten gewöhnlich ein kontrolliertes Vokabular und strukturieren damit spezielle Domänen, bspw. molekulare Funktionen für Proteine oder anatomische Strukturen verschiedener Spezies. Das Vokabular selbst enthält Konzepte, welche Entitäten der Domäne repräsentieren und durch Relationen, insbesondere is-a und part-of, in einem Graph oder Baum miteinander verbunden sind.

Durch die wachsende Popularität kommt dem Management von Ontologien eine immer wichtigere Rolle zu. Einerseits müssen für eine einheitliche Nutzung von Ontologien, z.B. in einem Grid, Heterogenitäten zwischen unterschiedlich entwickelten Ontologien aufgelöst werden. Hierzu zählen insbesondere die verschiedenen Formate und Speicherformen für Ontologien, z.B. relationale Datenbanken, XML, textbasierte Beschreibungen oder Standardisierungen wie OBO [6] in den Lebenswissenschaften. Auf der anderen Seite unterliegen Ontologien einer ständigen Weiterentwicklung (Evolution), da neue Forschungsergebnisse bzw. erweiterte Erkenntnisse in deren Entwicklung einfließen. Um die Evolution von Ontologien besser zu verstehen und deren Auswirkungen auf andere Strukturen (z.B. Annotationen, Ontologiemappings) zu ermitteln, erscheint es hilfreich Analysen über die Evolution von Ontologien durchzuführen. Aus diesen Gründen widmet sich dieser Beitrag zwei Arbeiten im Bereich Ontologiemanagement in den Lebenswissenschaften:

- Vorstellung einer Middleware zur Integration heterogener Ontologien in Gridumgebungen, insbesondere die Integration von Ontologien in MediGRID[1]
- Beschreibung eines Frameworks zur quantitativen Analyse von Ontologieevolution inklusive der vergleichenden Evolutionsanalyse von 16 Ontologien aus den Lebenswissenschaften

Der Rest des Beitrags gliedert sich wie folgt. Das zweite Kapitel beschreibt eine Middleware zur Integration von Ontologien in Grids und deren Anwendung in MediGRID. Kapitel 3 stellt ein Framework zur quantitativen Analyse von Ontologieevolution vor und berichtet über ermittelte Ergebnisse. Ein Ausblick auf weitere Arbeiten bildet den Abschluss des Beitrags.

[1] http://www.medigrid.de

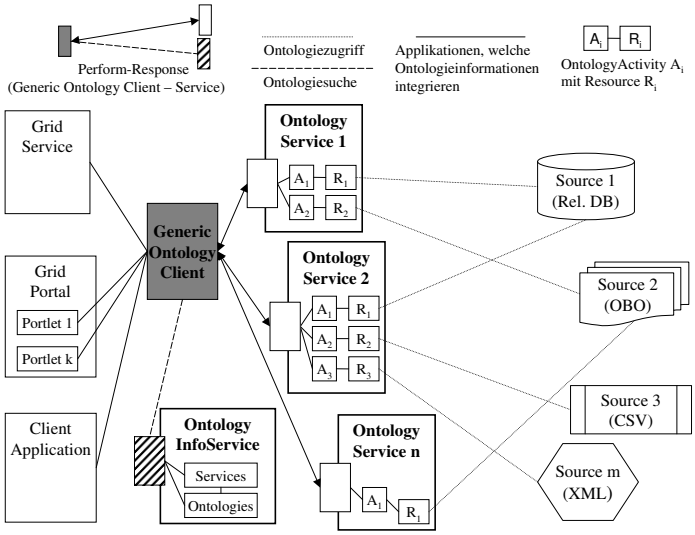

Abbildung 1: Architektur der Middleware

2 Integration von Ontologien in Grids

Um Ontologien verschiedener Formate und Herkunft einheitlich in Gridumgebungen integrieren zu können, wurde eine Middleware zur Integration von Ontologien in Grids entwickelt [3]. Die Middleware, sowie deren Nutzung in MediGRID werden in den beiden folgenden Abschnitten erläutert.

2.1 Architektur

Die in Abb. 1 dargestellte Architektur vermittelt einen Überblick über die Middleware, welche die folgenden Komponenten beinhaltet: (1) *Ontology Sources* als Ontologiequellen, (2) *Ontology Services* und der *OntologyInfoService* sowie (3) ein *Generic Ontology Client* als API für Gridanwendungen. Ontology Sources bezeichnen die ursprünglichen Ontologiequellen, welche Ontologieinformationen über Konzepte bzw. Beziehungen zwischen Konzepten enthalten. In den Lebenswissenschaften existieren die unterschiedlichsten Formate und Speicherformen für Ontologien. Bspw. wird die stark verbreitete GeneOntology (GO) [8] als relationale Datenbank veröffentlicht. Darüber hinaus existieren Standardisierungsbemühungen, wie z.B. das OBO-Format, um Ontologien in einem einheitlichen Format zu entwickeln. Weitere textbasierte Formate reichen von einfachen CSV Dateien bis zu den aus dem Semantic Web bekannten Formaten RDF bzw. OWL.

Im Kern der Middleware befinden sich Ontology Services, welche einen einheitlichen Zugriff auf eine bestimmte Anzahl zugeordneter Ontology Sources ermöglichen. Die Ontology Services basieren auf DataServices des OGSA-DAI Frameworks [4], welches in vielen Gridprojekten als Quasi-Standard für den Zugriff auf strukturierte Daten verwendet wird. Die Zuordnung von Ontology Services zu Ontology Sources ist $m : n$, d.h. ein Ontology Service kann auf mehrere unterschiedliche Ontology Sources zugreifen, umgekehrt kann eine Ontology Source von mehreren Ontology Services integriert werden.

Um diverse Ontologien einheitlich integrieren zu können, werden einerseits in jedem Ontology Service Ressourcenbeschreibungen für die zu integrierenden Ontologien verwaltet. Andererseits wird die im OGSA-DAI Framework verfügbare Activity Komponente verwendet, um spezielle Zugriffsroutinen (OntologyActivities) für die Integration von Ontologieinformationen zu realisieren. Die Kombination aus Ressourcenbeschreibung und OntologyActivity ermöglicht es einheitliche Gridschnittstellen für den Zugriff auf Ontologieinformationen bereitzustellen. Die Gridschnittstellen beinhalten typische Ontologieabfragen, wie z.B. Suche in Ontologien, Ermittlung von Konzeptattributen (Name, Definition, Synonyme, Referenzen) oder Ermittlung von Relationen zwischen Konzepten (Super- bzw. Subkonzepte).

Um die verteilten Ontology Services und ihre zugehörigen Ontology Sources im Grid lokalisieren zu können, wurde zudem ein zentraler OntologyInfoService realisiert. Der Service

dient vorzugsweise dem Discovery erreichbarer Ontologien im Grid und ist somit Ausgangspunkt für die Interaktion mit den Ontology Services. Die clientseitige Interaktion mit den jeweiligen Services übernimmt eine zentrale API, der Generic Ontology Client. Die Client API kann in unterschiedlichen Clients, z.B. Portlets in Portalen, Grid-Services im Grid oder eigenständigen Applikationen, eingesetzt werden. Die bereitgestellten Schnittstellen des Clients ermöglichen ein Interaktion, d.h. sie erlauben den Zugang zu Ontologieinformationen, ohne Kenntnis über technische Hintergründe, wie bspw. Speicherort einer Ontologie im Grid oder deren Format.

2.2 Anwendung in MediGRID

Die zuvor beschriebene Middleware wurde in MediGRID für die Integration diverser Ontologien aus den Lebenswissenschaften verwendet. MediGRID verwendet als zentralen Zugang zu seinem Grid ein Applikationsportal in welchem die unterschiedlichen Anwendungen (Bioinformatik, Bildverarbeitung, Klinische Forschung) einfach und web-basiert über Portlets erreichbar sind. Die Integration der Ontologien in das MediGRID Portal geschieht über zwei verschiedene Wege. Einerseits existiert ein zentraler Ontology LookUp Service, welcher Such- und Browsingfunktionalitäten für die integrierten Ontologien anbietet. Andererseits können MediGRID-Applikationen, wie bspw. AUGUSTUS [7] oder SNP Selection [1] ihre Daten unter Verwendung des Generic Ontology Clients mit Ontologieinformationen verknüpfen.

Aktuell wurden insgesamt 14 verschiedene Ontologien aus den Lebenswissenschaften mit Hilfe der Middleware in das MediGRID integriert. Hierzu zählen einerseits die stark verbreitete GeneOntology und der am National Cancer Institute (NCI) verwendete NCIThesaurus [5]. Weiterhin wurde mit RadLex ein Lexikon mit Begriffen aus dem radiologischen Bereich verfügbar gemacht. Weitere eher domänenspezifische Ontologien der OBO Initiative wurden ebenfalls integriert und werden in Anwendungen verwendet (z.B. SequenceOntology).

3 Quantitative Analyse der Evolution von Ontologien

Neben der Integration von Ontologien in bestimmte Umgebungen, hier am Beispiel Grid erläutert, spielt deren Evolution, d.h. deren Weiterentwicklung basierend auf neuen Erkenntnissen, im Ontologiemanagement eine wichtige Rolle. Die folgenden Abschnitte beschreiben Teile eines Frameworks zur quantitativen Evolutionsanalyse von Ontologien und angrenzenden Strukturen (z.B. Annotations- oder Ontologiemappings) [2]. Des weiteren werden Ergebnisse der Evolutionsanalyse für 16 Ontologien aus den Lebenswissenschaften präsentiert.

3.1 Ontologie- und Evolutionsmodell

Eine Ontologie in einer bestimmten Version v zu einem Zeitpunkt t wird durch folgendes Tupel charakterisiert: $ON_v = (C, R, t)$. Einerseits besteht eine Ontologie aus einer Menge von Konzepten C, welche über binäre Relationen R miteinander verbunden sind. C und R bilden gemeinsam mit den *roots* (Konzepte die keine Vorgänger besitzen) einen azyklisch gerichteten Graph (DAG). Konzepte selbst beinhalten diverse Attribute, in diesem Beitrag werden insbesondere zwei Attribute, das *accession* Attribut sowie die *obsolete* Information über den Status eines Konzepts verwendet. Das *accession* Attribut dient der eindeutigen Identifikation eines Ontologiekonzepts innerhalb einer Ontologie, während andererseits der Status eines Konzepts darüber informiert, ob ein Konzept aktiv ist (noObsolete) oder ob es veraltet ist und nicht mehr verwendet werden soll (obsolete).

Das Evolutionsmodell des Frameworks unterscheidet zwischen 3 grundlegenden Evolutionsoperatoren: *add*, *delete* und *toObsolete*. Während add die Einfügung neuer Objekte erfasst, werden durch delete Löschungen von Objekten beschrieben. Der insbesondere bei Ontologien angewendete toObsolete Operator erfasst Objekte, welche ihren Status von noObsolete auf obsolete ändern. Für die quantitative Evolutionsanalyse werden analog zu den Operatoren Evolutionsmengen berechnet. Als Eingabe dienen Objektmengen einer Version v_i (O_{vi}) und der geänderten Version v_j (O_{vj}) einer Quelle. Um veraltete von aktiven Objekten unterscheiden zu können, werden Submengen $O_{vi,obs}$ für veraltete und $O_{vi,nonObs}$ für aktive Objekte gebildet. Im Fall der Ontologieevolution können Objekte Konzepte C oder Relationen

R darstellen. Die Evolutionsmengen werden dann wie folgt berechnet: $add_{vi,vj} = O_{vj}/O_{vi}$, $del_{vi,vj} = O_{vi}/O_{vj}$ und $toObs_{vi,vj} = O_{vj,obs} \bigcap O_{vi,nonObs}$. Die Berechnung der Mengen erfolgt auf Basis eines Vergleichs von IDs (accessions) der beteiligten Objekte. Existiert bspw. ein Objekt in Version v_j und nicht in v_i, so wird dieses Objekt $add_{vi,vj}$ zugeordnet, umgekehrt würde das Objekt in $del_{vi,vj}$ eingefügt werden.

3.2 Metriken

Um die Evolution quantitativ erfassen zu können, werden Metriken definiert. Aus Platzgründen werden in diesem Beitrag nur einige relevante Metriken beschrieben.

Auf Basis der Objektmengen werden die folgenden Basismetriken definiert:

- $|O_{vi}|$: Anzahl von Objekten in Version v_i; $O \in$ {Konzepte C, Relationen R}
- $|C_{obs}|$, $|C_{nonObs}|$: Anzahl veralteter (obsolete) bzw. aktiver Konzepte

Aufbauend auf den Evolutionsmengen werden folgende Metriken zur Einschätzung der Evolution definiert:

- $Add_{vi,vj} = |add_{vi,vj}|$: Anzahl eingefügter Objekte zwischen v_i und v_j
- $Del_{vi,vj} = |del_{vi,vj}|$: Anzahl gelöschter Objekte zwischen v_i und v_j
- $Obs_{vi,vj} = |toObs_{vi,vj}|$: Anzahl zu obsolete markierter Objekte zwischen v_i und v_j
- $Add_{p,t}$, $Del_{p,t}$, $Obs_{p,t}$: durchschnittliche Anzahl veränderter Objekte pro Intervall t in einem Zeitraum p

Einerseits werden Versionsübergänge direkt quantifiziert, andererseits ermöglichen erweiterte Metriken die Analyse der Evolution innerhalb eines Zeitraums p für ein bestimmtes Änderungsintervall t. Bspw. können monatliche oder jährliche Änderungsraten (t) für einen gewissen Zeitraum (p) berechnet werden. Um nicht nur absolute Änderungen zu erfassen, werden auf deren Basis relative Änderungen wie auch eine add-delete ratio (adr) definiert: $adr_{vi,vj} = Add_{vi,vj}/(Del_{vi,vj} + Obs_{vi,vj})$, $add - frac_{vi,vj} = Add_{vi,vj}/|O_{vj}|$, $del - frac_{vj,vi} = Del_{vi,vj}/|O_{vi}|$, $obs - frac_{vj,vi} = Obs_{vi,vj}/|O_{vi}|$. Die relativen Änderungen sind ebenfalls auf beliebige Zeitintervalle und Perioden anwendbar: $add-frac_{p,t}$, $del-frac_{p,t}$ und $obs-frac_{p,t}$.

3.3 Ergebnisse der Evolutionsanalyse

Für die Analyse wurden 16 Ontologien mit insgesamt 386 Versionen in ein zentrales Repository integriert. Die verschiedenen Versionen stammen aus dem Zeitintervall [Mai 2004,Feb 2008], also einem Beobachtungszeitraum von 45 Monaten. Im Durchschnitt stieg die Anzahl von Ontologiekonzepten $|C|$ um 60%, wobei GeneOntology und NCIThesaurus ein Wachstum von 50% bzw. 78% aufweisen. Mithilfe der beschriebenen Metriken wurden die durchschnittlichen Änderungsraten pro Monat (absolut wie relativ) für den ganzen Beobachtungszeitraum

	Anzahl von Konzepten |C|		Mai 04 - Feb 08							Feb 07 - Feb 08		
Ontologie	*Start*	*Feb 08*	*Add*	*Del*	*Obs*	*adr*	*add-frac*	*del-frac*	*obs-frac*	*Add*	*Del*	*Obs*
NCI Thesaurus	35.814	63.924	627	2	12	42,4	1,3%	0,0%	0,0%	416	0	5
GeneOntology	17.368	25.995	200	12	4	12,2	0,9%	0,1%	0,0%	222	20	5
-- Biological Process	8.625	15.001	146	7	2	16,2	1,2%	0,1%	0,0%	133	10	2
-- Molecular Function	7.336	8.818	36	3	2	6,8	0,4%	0,0%	0,0%	69	7	3
-- Cellular Components	1.407	2.176	18	2	0	8,9	1,0%	0,1%	0,0%	19	3	0
ChemicalEntities	10.236	18.007	256	62	0	4,1	1,8%	0,5%	0,0%	384	67	0
FlyAnatomy	6.090	6.222	5	1	1	3,3	0,1%	0,0%	0,0%	6	0	0
MammalianPhenotype	4.175	6.077	65	2	9	6,0	1,2%	0,0%	0,2%	74	2	3
AdultMouseAnatomy	2.416	2.745	11	0	0	30,9	0,4%	0,0%	0,0%	1	0	0
ZebrafishAnatomy	1.389	2.172	33	5	1	5,5	1,8%	0,3%	0,1%	45	2	1
Sequence	981	1.463	19	3	2	4,1	1,5%	0,3%	0,2%	19	0	0
ProteinModification	1.074	1.128	5	2	1	1,5	0,4%	0,2%	0,1%	7	0	2
CellType	687	857	5	1	0	2,8	0,7%	0,2%	0,1%	1	0	0
PlantStructure	681	835	5	0	1	6,1	0,7%	0,0%	0,1%	3	0	0
ProteinProteinInteraction	194	819	21	0	0	41,7	2,7%	0,0%	0,2%	4	0	0
FlyBaseCV	658	693	1	0	1	2,1	0,2%	0,0%	0,1%	0	0	0
Pathway	427	593	7	1	0	7,9	1,3%	0,2%	0,0%	6	2	0

Abbildung 2: Änderungen in analysierten Ontologien (sortiert nach $|C|$ absteigend)

als auch für das letzte Jahr ermittelt (Abb. 2). Die Analyse ergab, dass die großen Ontologien ($|C| > 10000$, dunkelgrau hinterlegt) ebenfalls die größten Änderungsraten erfahren: 360 (25) Einfügungen (Löschungen) pro Monat im Vergleich zu 86 (6) Einfügungen (Löschungen) in allen untersuchten Ontologien. Weiterhin kann man erkennen, dass Einfügungen in allen Ontologien dominieren, jedoch existieren einige Ontologien, z.B. ChemicalEntities, die ebenfalls hohe Löschraten bzw. obsolete Markierungen aufweisen. Die *adr* Metrik zeigt, dass bspw. NCIThesaurus im Schnitt 42 mal mehr hinzufügt als löscht bzw. obsolete markiert. Hingegen weist ChemicalEntities hier einen Wert von 4 auf, was im Durchschnitt einen Anteil von 20% Löschungen an den Gesamtänderungen bedeutet. Die relativen Änderungsraten deuten an, dass einige mittelgroße bzw. kleine Ontologien in Bezug auf ihre Größe durchaus hohe Evolutionsraten besitzen, z.B. ZebrafishAnatomy und ProteinProteinInteraction.

Interessant erscheint die Verwendung von obsolete zur Markierung veralteter Konzepte in den analysierten Ontologien. Während wenige Ontologien, z.B. ChemicalEntities und AdultMouseAnatomy, auf obsolete Markierungen komplett verzichten, nutzen 13 der 16 Ontologien einen hybriden Ansatz (gleichzeitige Löschungen und obsolete Markierungen). Eine weitere Analyse vergleicht die Evolution im gesamten Beobachtungszeitraum mit der im letzten Jahr. Hier zeigen sich 3 verschiedene Szenarien. Einerseits existieren Ontologien, welche in beiden Zeiträumen hohe Änderungsraten aufweisen: NCIThesaurus oder GeneOntology. Andere Ontologien wiederum wurden im letzten Jahr verstärkt weiterentwickelt, d.h. die Domänen bzw. Forschungsfelder dieser Ontologien erscheinen derzeit hoch aktuell, bspw. ChemicalEntities oder GO MolecularFunction. Die letzte Gruppe umfasst Ontologien, welche im letzten Jahr gegenüber dem Gesamtzeitraum keine bzw. wenige Änderungen aufzeigen. Hierzu zählen unter anderem AdultMouseAnatomy oder CellTypeOntology.

4 Zusammenfassung und Ausblick

Dieser Beitrag berichtete anhand zweier ausgewählter Beispiele über das Ontologiemanagement in den Lebenswissenschaften: (1) Integration heterogener Ontologien in Grids und (2) Quantitative Evolutionsanalyse von Ontologien. Möglichkeiten für künftige Arbeiten umfassen bspw. den Ausbau des Frameworks zur Evolutionsanalyse um weitere Evolutionstypen (z.B. Änderung an Attributen). Weiterhin ist die Verwaltung der unterschiedlichen Ontologieversionen im zentralen Repository effizienter gestaltbar. Zudem könnten Analyseergebnisse der Öffentlichkeit (z.B. Ontologiedesignern) zugänglich gemacht werden.

Acknowledgements Diese Arbeit wurde vom BMBF im Rahmen des Projektes “MediGRID - Ressourcenfusion für Medizin und Lebenswissenschaften” (01AK803E) gefördert.

Literatur

[1] J. Hampe, S. Schreiber, and M. Krawczak. Entropy-based SNP selection for genetic association studies. *Human Genetics*, 144:36–43, 2003.

[2] M. Hartung, T. Kirsten, and E. Rahm. Analyzing the Evolution of Life Science Ontologies and Mappings. *To appear in Proc. of DILS 2008*, 2008.

[3] M. Hartung and E. Rahm. A Grid Middleware for Ontology Access. *1st German eScience Conference*, 2007.

[4] K. Karasavvas et al. Introduction to OGSA-DAI Services. In *LNCS*, volume 3458, pages 1–12, 2005.

[5] N. Sioutos et al. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Web Semantics*, 40:30–43, 2007.

[6] B. Smith et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25:1251–1255, 2007.

[7] M. Stanke et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res*, 34(Web Server issue):435–439, 2006.

[8] The Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32:258–261, 2004.