String-Matching Basierter Vergleich Biomedizinischer Ontologien

Exposé zur Studienarbeit

Eingereicht von: Jonathan Bräuer
Geboren: 13.10.1988 in Herrenberg
Gutachter: Prof. Dr. Ulf Leser
Eingereicht am: 11.09.2015

1 Einführung

Um die automatisierte Verarbeitung biomedizinischer Daten zu erleichtern, haben sich in den vergangenen Jahren Ontologien als Werkzeuge durchgesetzt. Die einzelnen Ontologien wurden für einen bestimmten Zweck erstellt und beschreiben dadurch meist sehr spezielle Domänen. Es gibt Ontologien für die Phänotypen vieler Spezies, für Krankheiten, Taxonomien, Anatomien und viele weitere Bereiche. Natürlich unterscheiden sich diese voneinander, die verwendeten Domänen überschneiden sich aber oftmals. Es kann davon ausgegangen werden, dass viele der Konzepte also in mehr als einer Ontologie beschrieben werden. Damit sind Mappings zwischen Ontologien möglich, die einen Informationstransfer ermöglichen.

Mit der *Open Biological and Biomedical Ontologies Foundry (OBO)* hat sich ein Zusammenschluss von Forschern gebildet, in der unterschiedlichste Ontologien gesammelt werden. Die meisten Ontologien enthalten außerdem Verweise in externe Vokabulare, wie das *Unified Medical Language System (UMLS)*. In der Humanmedizin haben sich vor allem die *Human Disease Ontology (DO)* und die *Human Phenotype Ontology (HPO)* als Standards für die Forschung herausgebildet. Wie sich bereits an den Namen erkennen lässt, unterscheiden sich die Ansätze dabei. Während die Struktur der HPO sich aus den Phänotypen ergibt, liegt der Fokus der DO auf den bekannten Krankheiten. Da Krankheiten und Phänotypen jedoch eng verwand sind, ergibt sich die Frage, wie ähnlich sich diese Ontologien sind. Es existieren bereits viele Abbildungen zwischen den unterschiedlichsten Ontologien, zwischen diesen beiden gibt es aber keine mir bekannten Abbildungen.

In meiner Studienarbeit werde ich diese beiden Ontologien miteinander Vergleichen und bestehende Gemeinsamkeiten und Unterschiede mithilfe eines Mappings aufzeigen. Dabei werde ich zunächst bereits bestehende Arbeiten vorstellen und ein Verfahren entwickeln, mit dem die Strukturen der Ontologien miteinander verglichen werden sollen. Eine Menge von Abbildungen von einer Ontologie auf die andere soll erstellt werden, wobei jedes Konzept dabei auf höchstens ein Konzept in der anderen Ontologie abgebildet werden soll. Ein Konfidenzwert gibt dabei die vermutete Qualität jeder einzelnen Abbildung an.

Lässt sich ein Mapping finden, so könnten Forschungsergebnisse besser miteinander vernetzt werden. Lassen sich große Teilmengen der Ontologien finden, für die kein Mapping möglich ist, lässt sich erkennen, welche Ontolgie für einen bestimmten Zweck besser geeignet ist.

1.1 Biomedizinische Ontologien

Ontologien beschreiben im allgemeinen Typen von Entitäten und ihre Beziehung zueinander. Die hier verwendeten Ontologien setzten sich aus Konzepten, Metainformationen, welche die Konzepte genauer beschreiben (Annotationen) und binären Relationen zusammen. Die wichtigste Relation stellt dabei die 'is_a' Beziehung dar, welche die Verfeinerungen der Konzepte beschreibt. Da ein Konzept mehrerer Konzepte verfeinern kann entsteht ein gerichteter, azyklischer Graph. Jedes Konzept ist mit einem 'label' und gegebenenfalls 'has_synonym' Annotationen versehen, welche die Bezeichnungen für dieses Konzept beschreiben und durch 'has_DBXRef' in externen Quellen genauer definiert. Die Relationen können auch auf externe Ontologien oder Datenbanken verweisen. Beide Ontologien liegen im '.owl' (*Ontology Web Language*) Format vor und können mit der *OWLAPI (http://owlapi.sourceforge.net)* interpretiert werden.

Human Phenotype Ontology

Die *Human Phenotype Ontology (HPO)* wurde dafür geschaffen, das Vokabular phänotypischer Auffälligkeiten zu standardisieren. Es beinhaltet über 11.000 Terme, über 15.000 'is_a' Relationen und ist stark vernetzt mit anderen Ontologien (etwa 46% der Konzepte besitzen Referenzen zu anderen Ontologien) [**?**]. Außerdem sind die Konzepte mit logischen Definitionen versehen, wodurch die Konzepte durch Ausdrücke ergänzt werden, die Ort des Auftretens und Veränderung beschreiben. Diese Definitionen setzen sich sich dabei wiederum aus Verweisen auf andere Ontologien zusammen. *UBERON* ist eine Ontologie der Anatomie vieler unterschiedlicher Spezies und in der *Phenotype, Attribute and Trait Ontology (PATO)* werden phänotype Qualitäten beschrieben. Damit lässt sich zum Beispiel Hypoglykämie durch den Ausdruck:

'$< decreased\ concentration > towards < glucose > inhering\_in < blood >$'

beschreiben. Dies erleichtert die gleichzeitige Verwendung mehrerer Ontologien und vereinfacht es, Forschungsergebnisse unterschiedlicher Spezies zusammen zu bringen [**?**]. Die Ontologie ist frei Verfügbar unter *http://human-phenotype-ontology.org/*.

Human Disease Ontology

Die *Human Disease Ontology (DO)* bietet eine Sammlung standardisierter Bezeichnungen und Beschreibungen der Krankheiten, die Menschen befallen können. Dadurch wird es möglich, Daten aus unterschiedlichsten Domänen zu verknüpfen. Die DO beinhaltet über 6000 Konzepte und fast 7000 'is_a' Relationen und ist sehr gut zu externen Quellen vernetzt [**?**]. Auch die DO ist frei Verfügbar unter *http://disease-ontology.org*.

Unified Medical Language System

Das *Unified Medical Language System (UMLS)* ist zu einer der wichtigsten Vokabular- und Werkzeugsammlung in der Biomedizin geworden. Das System setzt sich aus einzelnen Komponenten zusammen, die eine Generalisierung der biomedizinischen Sprache erleichtern sollen. Die Komponente *Specialist* ist ein syntaktisches Lexikon mit einer Sammlung von Werkzeugen, das die automatisierte Verarbeitung unstrukturierter biomedizinischer Texte verbessern soll. Der *Metathesaurus* bietet eine mehrsprachige Datenbank von Vokabularen aus den Lebenswissenschaften und den Beziehungen zwischen deren Konzepten [**?**].

In dieser Arbeit sollen die genannten Komponenten von *UMLS* verwendet werden um ein Mapping zwischen der DO und der HPO zu ermöglichen, auch wenn semantische Unterschiede in den Bezeichnungen der Konzepte vorliegen.

Abbildung 1: Vergleich der Ontologien

Ontologie	Konzepte	*is_a*-Relationen	DB-Verweise	Bezeichnungen
HPO	11416	15249	14913	19041
DO	6598	6943	45887	17878

2 Stand der Forschung

2.1 Ontology Mapping

Ziel des *Ontology Mapping's* ist es, eine Abbildung zwischen Zwei Ontologien zu finden. Diese ist eine Menge von Konzeptpaaren, mit jeweils einem Konzept aus beiden Onto-

logien und Metainformationen, welche das Paar genauer beschreiben (zB. Konfidenz der Abbildung). Dabei kann es unterschiedliche Kardinalitäten geben, da ein Konzept sowohl auf ein einzelnes oder mehrere Konzepte der jeweils anderen Ontologie abgebildet werden kann (1:1, 1:n, n:1, n:m).

Solche Abbildungen können nützlich sein, um zum Beispiel unterschiedliche Versionen von Ontologien zu vergleichen, mehrere Ontologien ineinander zu integrieren oder Informationen zwischen Ontologien zu übertragen.

Moderne Systeme zum *Ontology Mapping* berücksichtigen dabei zum Beispiel die Terminologie (vorkommende Texte), die Struktur, bekannte Instanzen der Konzepte oder Aussagenlogische Terme der Ontologien, um eine möglichst gute Abbildung zu finden. Eine Übersicht über solche System bieten *Pavel Shvaiko et. al* [**?**].

Die *Ontology Alignment Evaluation Initiative (http://oaei.ontologymatching.org)* hilft dabei, diese Methoden zu evaluieren und die Systeme durch jährlich stattfindende Konferenzen und Wettbewerbe zu verbessern.

In dieser Arbeit wird eine einfache Methode zum Matching einzelner Konzepte mithilfe der Terminologie verwendet. Komplexere Methoden könnten bessere Ergebnisse oder eine Verifikation bieten, würden allerdings den Umfang dieser Arbeit überschreiten.

2.2 Mapping biomedizinischer Ontologien

Eine Methode für das Mapping biomedizinischer Ontologien bietet der *Lexical OWL Ontology Matcher (LOOM)*. Dabei werden alle Bezeichnungen und Synonyme der Konzepte zweier Ontologien paarweise miteinander verglichen. Können so sehr ähnliche Zeichenketten in beiden Konzepten aus jeweils einer Ontologie gefunden werden, so werden diese beiden Konzepte als Match angesehen. Dafür werden zuerst Trennzeichen (Leerzeichen, Unterstriche, Punkte...) entfernt. Diese resultierenden Zeichenketten dürfen sich bei einem Match in höchstens einem Zeichen unterscheiden (bzw. in keinem, wenn sie kürzer als 4 Zeichen sind). Diese Methode stellte sich trotz ihrer Einfachheit als effektiv heraus und die Ergebnisse sind mit denen komplexerer Methoden vergleichbar [**?**].

Um Ontologien von Phänotypen allgemein zugänglicher zu machen, wurde wie bereits erwähnt ein System entwickelt, das Phänotypen mithilfe von speziesübergreifenden Ontologien beschreiben soll. Dabei werden logische Relationen verwendet, um die Konzepte einer Ontologie allgemein zu beschreiben. Dies wird durch sogenannte *EQ*-Aussagen ermöglicht, wobei Objekte (E) mit Eigenschaften (Q) in Relation gestellt werden. Für diese *EQ*-Definitionen werden Verweise auf allgemeine, übergeordnete Ontologien verwendet. Die Beschreibungen der Eigenschaften sind dabei zum Beispiel in *PATO* beschrieben und die *UBER* Ontologie beschreibt die Objekte der Anatomie von Lebewesen. *EQ*-Definitionen können dabei noch durch weitere Restriktionen ergänzt werden, um den Phänotypen genauer zu beschreiben. Damit ist es möglich, Konzepte einer beliebigen Ontologie durch logische Relationen mit Konzepte aus allgemeineren Ontologien zu beschreiben.

PhenomeNET hat mit Hilfe dieser Aussagen eine kombinierte Ontologie mit 275.000 Konzepten der Phänotypen aus 5 Ontologien erzeugt (Mensch, Säugetier, Wurm, Hefe und Fliege). Mithilfe der *EQ*-Aussagen konnten äquivalente Konzepte (Matches) gefunden und mit einem OWL Reasoner auf Konsistenz überprüft werden [**?**].

Eine Eingliederung der *DO* in dieses System ist nicht mit demselben Verfahren möglich, da die Konzepte nicht mit *EQ*-Aussagen annotiert sind und es sich nicht bei allen Konzepten um Phänotypen handelt. Können jedoch die Konzepte der *DO*, welche Phänotypen beschreiben, gefunden werden, so ist eine Eingliederung in solche kombinierten Ontologien möglich.

3 Verfahren

Für die Abbildung der Ontologien werden folgende Schritte befolgt. Dabei soll das Verfahren allgemein gehalten werden, um auch für den Vergleich anderer Ontologien angewendet werden zu können.

3.1 Erweiterung der Synonyme und Verweise

In den Ontologien sind bereits viele Synonyme und Verweise auf externe Datenbanken definiert (Siehe Abbildung **??**). Sowohl die DO als auch die HPO enthalten vor allem Verweise auf UMLS, die DO enthält unter anderem weitere Verweise zu ICD Codes, NCI Thesaurus, und OMIM. Die HPO enthält Verweise zu ICD Codes, Orphanet, DECIPHER und OMIM. Allerdings sind nicht alle Konzepte mit Verweisen annotiert und es gibt nur wenige Verweise, die sowohl in Konzepten der DO und HPO verwendet werden.

Es sollen weitere Verweise auf *UMLS Metathesaurus* gefunden werden, da dieses System ein sehr umfassendes Vokabular enthält. Mithilfe dieser gefundenen und den bereits in den Ontologien definierten Verweisen sollen dann die Synonymlisten erweitert werden. Da alle externen Quellen frei zugänglich sind, können diese einfach über ein Web Interface oder über eine lokale Installation ausgelesen werden.

3.2 Normalisierung

Um das Matching zu vereinfachen werden die Bezeichungen normalisiert. Unnötige Trennzeichen werden entfernt, alle Großbuchstaben werden in Kleinbuchstaben umgewandelt. Die gegebenen externen Verweise müssen überprüft werden, da diese sich teils auf unterschiedliche Versionen der Quellen beziehen. Außerdem soll überprüft werden, ob mithilfe von *UMLS Specialist* Teilterme der Bezeichnungen durch ein präferiertes Synonym ersetzt werden können, um semantische Unterschiede in den Bezeichnungen zu umgehen.

3.3 Matching

Für das Matching werden alle Konzepte beider Ontologien paarweise miteinander verglichen, indem für alle Bezeichnungen ('label' und 'has_synonym') beider Konzepte mit einem String-Matching Verfahren eine Konfidenz für ein Match angegeben wird. Diese Konfidenz errechnet sich dabei daraus, wie ähnlich sich die Mengen der Bezeichnungen beider Konzepte sind und woher die Bezeichnungen bekannt sind (in der Ontologie definiert oder in Schritt 3.1 ergänzt). Welches Verfahren für dieses String-Matching verwendet werden soll, steht zu diesem Zeitpunkt noch nicht fest. In einer Arbeit von *Sun et al.* ([**?**]) wurden verschiedene String-Matching Algorithmen im Kontext von Ontology Mapping verglichen. Diese dienen als Ansatzpunkt.

Abbildung 2: Beispiel für Überschneidungen in Matches

Ontologie	Relation	Label	ID
DO	|	dry eye syndrome	[DOID_10140]
	← is_a	keratoconjunctivitis sicca	[DOID_12895]
	← is_a	xerophthalmia	[DOID_10138]
HPO	|	Keratoconjunctivitis sicca	[HP_0001097]
	← has_synonym	Xerophthalmia	[HP_0001097]

3.4 Validierung der Konsistenz

Problematisch bei diesem Vorgehen sind vermutlich unterschiedliche Granularitäten und möglicherweise mehrfaches Vorkommen gleicher oder sehr ähnlicher Bezeichnungen in den Ontologien, was zu falsch erkannten Matches führen kann. In Abbildung **??** ist ein Beispiel für eine Mehrfacherkennung. Die Konzepte 'keratoconjunctivitis sicca' und 'xerophthalmia' in der DO teilen sich beide eine Bezeichung mit der Klasse 'Keratoconjunctivitis sicca' in HPO, da dieses als Synonym ebenfalls 'Xerophthalmia' besitzt. Aus der Struktur lässt sich erkennen, das die Konzepte in der HPO in diesem Fall in einer gröberen Granularität vorliegen, da beide Konzepte in der DO dem Konzept 'dry eye syndrome' zugehörig sind.

Es müssen also mögliche Muster in der Struktur erkannt werden, welche ein zulässiges (teilweises) Mapping darstellen. Lassen sich bei Mehrfacherkennung keine solchen Muster finden, müsste das Mapping manuell entschieden werden. Um diese Nachbearbeitung zu vermeiden, werden für diese teilweisen Matchings die Konfidenzwerte angepasst.

3.5 Evaluierung des Verfahrens

Da für das String-Matching ein bereits bekannter Algorithmus verwendet werden soll, wird auf eine Evaluierung dieses Algorithmus verzichtet.

Um zu überprüfen, ob sich durch die Erweiterung der Synonyme (Schritt 3.1) und die Normalisierung (Schritt 3.2) ein Informationsgewinn erkennen lässt, sollen die jeweils gefundenen Mappings miteinander verglichen werden. Die *Precision* des Systems soll durch eine zufällig gewählte, kleine Teilmenge des Mappings manuell abgeschätzt werden. *Recall* lässt sich auf diese Weise nicht bestimmen, da es zuviele mögliche Mappings gibt, um alle zu überprüfen. Eine Möglichkeit, um das System zu evaluieren, wäre die Anwendung auf Ontologien, mit bereits bekannten Mappings. Dafür würde sich ein Mapping von der HPO auf die *Mammalian Phenotype Ontology* eignen, da dieses im Rahmen von *PhenomeNET* vorliegt [**?**]. Es muss überprüft werden, ob sich dieses Mapping für eine Evaluierung eignet, oder ob es andere Ontologien mit Mappings gibt, der dafür verwendet werden könnten.

Literatur

[1] *The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data*
S. Köhler et al.
Nucleic Acids Research 2014 Jan;42(Database issue):D966-74

[2] *Disease Ontology: a backbone for disease semantic integration*
L. M. Schriml, C. Arze, S. Nadendla, Y. W. Chang, M. Mazaitis, V. Felix, G. Feng, W. A. Kibbe
Nucleic Acids Research 2012 40 (D1): D940-D946.

[3] *Integrating phenotype ontologies across multiple species*
C. J. Mungall, G. V. Gkoutos, C. L. Smith, M. A. Haendel, S. E. Lewis, M. Ashburner
Genome Biology 2010, 11:R2

[4] *Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research*
S. Köhler, S. C. Doelken, B. J. Ruef, S. Bauer, N. Washington, M. Westerfield, G. Gkoutos, P. Schofield, D. Smedley, S. E Lewis, P. N. Robinson, C. J. Mungall
F1000Research 2014, 2:30

[5] *A Comparative Evaluation of String Similarity Metrics for Ontology Alignment*
Y. Sun, L. Ma, S. Wang
Journal of Information & Computational Science 12:3 (2015)

[6] *Creating mappings for ontologies in biomedicine: simple methods work*
A. Ghazvinian, N. F. Noy, M. A. Musen
AMIA Annu Symp Proc 2009, 2009:198-202.

[7] *Quantitative comparison of mapping methods between Human and Mammalian Phenotype Ontology*
A. Oellrich, G. V. Gkoutos, R. Hoehndorf, . Rebholz-Schuhmann
Ontologies in Biomedicine and Life Sciences, Berlin, 2011

[9] *PhenomeNET: a whole-phenome approach to disease gene discovery.*
R. Hoehndorf, P. N. Schofield, G. V. Gkoutos
Nucleid Acids Research, 2011

[10] *Ontology matching: state of the art and future challenges*
P. Shvaiko, J. Euzenat
IEEE Transactions on Knowledge and Data Engineering, 2013

[11] *UMLS® Reference Manual [Internet].*
Bethesda (MD): National Library of Medicine (US); 2009 Sep-. 1, Introduction to the UMLS.
http://www.ncbi.nlm.nih.gov/books/NBK9675/