Bio-Ontologien: Darstellung in GO und OWL

1. Kurzeinstieg

Ontologien:
- genau Begriffsdefinition ist schwierig
- Wissensrepräsentation eines formal definierten Systems von Begriffen und Relationen
- stellt Sein-Zusammenhänge eines Weltausschnittes dar
- zusätzlich mit Regeln versehen, so dass Schlussfolgerungen möglich sind

Wissensrepräsentationssprachen:
- Semantik macht Wissen sowohl für Mensch, als auch Maschine erfassbar
- es existieren verschiedene Sprachen
  - bieten unterschiedliche Möglichkeiten Wissen zu formalisieren
  - ermöglichen verschiedene Anfrage- und Auflösungstechniken
  - es gibt keine ultimative Sprache, alle haben Vor- und Nachteile
  - ➲ Übersetzung zwischen Sprachen ist erstrebenswert um möglichst viel Nutzen zu können
  - ABER: Sprachen sind nicht gleich mächtig
    - enthaltenes Wissen muss angepasst werden
    - Änderungen sollte dabei so gering wie möglich ausfallen
    - ➲ explizites Verständnis der Semantik notwendig

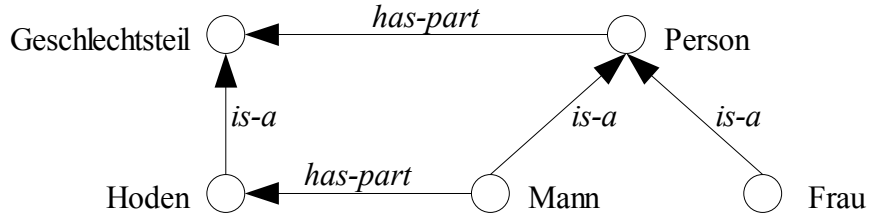
Reasoner (logisch Denkender):
- Programm das Anfragen basierend auf einer Ontologie beantwortet

OWL:
- Vom W3C empfohlene Web Ontology Language zur Darstellung von Wissen
- OWL-DL Reasoner ermöglichen automatische und manuelle Überprüfung des Wissens
  - kann Konsistenz der Wissensbasis prüfen und auf Klassenhierarchie abbilden

Go (Gene Ontology):
- de facto Standard zur Beschreibung von Genprodukten wie
  - Molekularfunktionen
  - Zellbestandteilen
  - biologischen Prozessen
- Teil des „Umbrella Projektes“ und damit der „Open Biomedical Ontologies“ (OBO)
- verwendet sowohl Terme zur Beschreibung als auch natürlich sprachliche Interpretation dieser
- Wissensrepräsentation basiert auf einem DAG (Directed Acyclic Graph)
  - in verschiedenen Formaten darstellbar (u.a. MySql, XML, OWL, OBO)

2. Warum ist die Semantik so wichtig?

Beispiel:



- Darstellung scheint zunächst eindeutig, aber:
  - Sind alle Männer Personen?
  - Kann eine Person nur Mann oder Frau sein?
  - Kann eine Person gleichzeitig Mann und Frau sein?
  - Wie viele Hoden hat ein Mann?
  - Bedeutet Hoden haben, dass es sich um einen Mann handelt?
  - Haben alle Männer Hoden?
  - Sind Hoden die einzigen Geschlechtsteile von Männern?
  - Sind Hoden immer Teil eines Mannes?
- Gründe:
  - Menschen haben Vorwissen und abstrahieren davon
  - man vermutet, legt nahe, rät, oder weiß die Antwort - oder eben nicht
  - für Computer sind all das nur Symbole, er hat keinerlei Vorwissen
- zu glauben die Bedeutung einer Ontologie zu verstehen kann gefährlich sein
  - Mehrdeutigkeiten werden ggf. nicht erkannt (weil sie oft trivial scheinen)
  - erschwert die Einordnung der Ergebnisse (richtig, falsch, vollständig)
- ➲ genaue und eindeutige Definition notwendig damit durch Computer korrekt erfassbar
- ➲ präzise Angaben nötig damit Ergebnisse durch Menschen überprüfbar und nachvollziehbar

3. OWL

- OWL-DL ist eine Ontologie-basierte Beschreibungssprache (Description Language)
- Formalisierungen durch „Objekte“, „Klassen“, „Subklassen“ und ihren „Beziehungen“
- besteht im Wesentlichen aus Fragmenten der Prädikatenlogik erster Ordnung
- sehr ausdrucksstark, verwendbare Attribute sind u.a.
  - disjoint: Klassen sind disjunkt (z.B. Mann und Frau)
  - someValueFrom: teil-von-Beziehung von existenzieller Bedeutung
  - allValuesFrom: alle teil-von-Beziehungen müssen aus bestimmter Menge kommen
  - complementOf: logisches not
  - IntersectionOf: Schnittmenge
  - complete: Bedingung ist notwendig und hinreichend (z.B. Mann <--> Hoden)
  - partiell: Bedingung ist Notwendig, aber nicht hinreichend (z.B. Auto --> Rad)

Umsetzung des obigen Beispiels in OWL-DL:

DisjointClasses(**Mann Frau**)
SubClassOf(**Person** (UnionOf(**Mann Frau**)
Class(**Mann** compete IntersectionOf(**Person** Restriction(has-part someValueFrom **Hoden**)))
Paraphrase: Männer sind alle Personen die, neben anderen Dingen, Hoden haben.

Class(**Frau** complete IntersectionOf(**Person** Restriction(has-part allValuesFrom complementOf(**Hoden**))))
Paraphrase: Frauen sind alle Personen, die neben anderen Dingen, keine Hoden haben.

Problem:

- wir wissen nichts über Hoden, außer dass Männer sie besitzen
  - männliche Instanzen anderer Wesen müssen keine Hoden besitzen
  - Hoden können auch als Teil anderer Objekte auftreten
  - oder sogar Teil von nichts sein, also für sich allein existieren
- ➲ um „heimatlose“ Hoden zu vermeiden definieren wir sie besser als Teil von Tieren
- ➲ außerdem muss definiert werden, dass has-part tatsächlich invers zu part-of ist

Zusätzliche Beschreibung von Hoden:

Objektproperty(has-part InverseOf part-of)
Class(**Hoden** partial IntersectionOf(**Organ** Restriction(part-of someValueFrom Intersection(**Male Animal**))))
Paraphrase: Hoden sind Organe die nur in männlichen Tieren auftreten.

Hinzufügen von Eunuchen:

Class(**Eunuch** partial IntersectionOf(**Person** complementOf(Restriction(has-part someValueFrom **Hoden**))))
Paraphrase: Eunuchen sind alle Personen, die neben anderen Dingen, keine Hoden haben.

Problem:

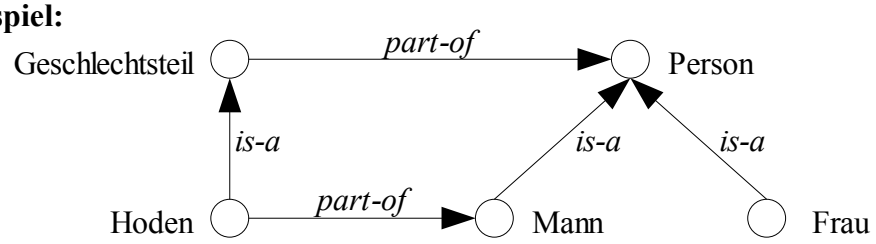
- Überprüfung mit einem Reasoner ergibt: Eunuchen sind Subklasse von Frauen
- wenn wir sie jedoch als Subklasse von Männern definieren ist die Ontologie inkonsistent
  - es kann keine Instanz von ihnen auftreten
  - ➲ Instanz von Mann ist schlecht definiert, Modell muss überarbeitet werden
- OWL-DL erlaubt es Wissen darzustellen und die korrekte Interpretation zu überprüfen

Bemerkung:
OWL ist wesentlich mächtiger. Es wurde nur ein kleiner Ausschnitt gezeigt.
Mehr Informationen unter: http://www.w3.org/TR/owl-guide

4. GO

- Semantik basiert im Gegensatz zu OWL nicht auf logischer Formalisierung
- Wissen wird als DAG (Directed Acyclic Graph) über Knoten und Kanten dargestellt
- Beschriftung benennt symbolisierte Klasse (Knoten) bzw. Art der Beziehungen (Kanten)
- Kanten können nur mit part-of oder is-a beschriftet werden
- Kanten sind immer vorwärts gerichtet (da Graph azyklisch)
- Kanten gelten immer für einen Knoten und all seine Kinder
- Knoten können weitere Informationen enthalten:
  - Synonymangaben: exact, broad, narrow, related
  - Beschreibung in natürlicher Sprache ➲ Gefahr von Mehrdeutigkeit
  - Randinformationen wie Autor, Quelle, etc.
- ➲ Einsatz automatisierter Reasoner zur Deduktion ist nicht beabsichtigt

Beispiel:



Achtung:
- verweiste Knoten sollten vermieden werden (mit part-of aber nicht is-a Beziehung versehen)
- sie deuten auf unpräzise Modellierung hin

Probleme:
- natürlich sprachliche Beschreibung ist oft nicht eindeutig
- keine Möglichkeit Eltern/Kind-Beziehungen als disjunkt oder überlappend festzulegen
- die Zuordnungen part-of ist nicht klar als hinreichen und/oder notwendig definierbar
  - A part-of B interpretierbar als:
    1. A kann Teil von B sein, oder nicht; B kann A enthalten, oder nicht
    2. A ist immer Teil von B
    3. B beinhaltet immer A
    4. A ist immer Teil von B und B beinhaltet immer A (2. + 3.)

„True Path Regel“ als Lösungsansatz für Eindeutigkeit:
- Der Pfad von einem Kind zu seinen top-level Eltern (is-a oder part-of) muss immer wahr sein.
- ➲ jede Instanz einer Klasse muss immer Instanz ihrer Superklasse(n) sein (Transitivität)
- part-of-Interpretation 1 und 3 verletzten dieses Gesetzt
- 4 wird vom Style Guide nicht empfohlen da er der nicht vorwärts gerichtet ist
- ➲ nur noch Interpretation 2 „A ist immer Teil von B“ möglich
  - ➲ Verlust an Ausdruckskraft: wir können nichts über B's Zugehörigkeit sagen

5. Übersetzung

GO DAG zu OWL-DL:
- Disjunktheit und Überlappung von Kindern kann nicht automatisch ermittelt werden
- ➲ automatische Übersetzung nur unter der Annahme, dass alle Klassen sich überlappen können
- part-of Interpretation 2 sollte angenommen werden und kann ggf. manuell zu 4 ergänzt werden
- ➲ automatische Übersetzung unter Annahmen möglich und sinnvoll

OWL-DL zu GO DAG:
- Synonymklassen werden entweder zu einer vereinigt (exact), oder als Subklassen dargestellt
- Überlappungseigenschaft von Kindern nicht darstellbar
- ➲ ggf. verlustbehaftete Übersetzung
- Entstehender Graph muss nicht zwingend azyklisch sein, aber:
  - reine is-a Zyklen können zu einer einzigen Klasse verschmolzen werden
  - part-of Zyklen sind anti-intuitiv
  - Kantenbeschriftung in OWL-DL frei, daher sind doch Zyklen möglich (z.B. interagiert-mit)
- GO unterstützt nur part-of und is-a Beziehungen, alles darüber hinaus ist ggf. nicht darstellbar
- ➲ automatische Übersetzung nur im Idealfall möglich und eventuell verlustbehaftet

6. Zusammenfassung

- GO soll gemeinschaftliches Domänen-Wissen für alle verfügbar machen
- Mann-Frau Beispiel zeigt aber wie schnell Mehrdeutigkeit vorliegt
  - dadurch entstehen schnell Interpretationsprobleme
  - ➲ Ontologie nicht automatisch auswertbar
  - ➲ selbst manuelles Lesen bei komplexeren Ontologien ggf. schwer möglich
- Umwandlung von GO DAG nach OWL-DL ist möglich
  - bietet Verfeinerungsmöglichkeiten um Beziehungen genauer zu machen
  - ermöglicht Maschinenlesbarkeit und automatische Auswertung
  - dadurch ist Wissen besser überprüfbar (Inkonsistenz-Detektion, Klassenhierarchie)
  - on-the-fly kann neues, nicht in der Ontologie erfasstes Wissen eingebunden werden

7. Quellen

Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL [http://www.biomedcentral.com/content/pdf/1471-2105-8-57.pdf]

OWL Web Ontology Language Guide [http://www.w3.org/TR/owl-guide]

Open Biomedical Ontology [http://obo.sourceforge.net]

The GO Editorial Style Guide [http://www.geneontology.org/GO.usage.shtml]