

Aus dem  
Max-Delbrück-Zentrum für molekulare Medizin (MDC)  
Berlin-Buch

DISSERTATION

**Ein Repräsentationsformat zur  
standardisierten Beschreibung und  
wissensbasierten Modellierung genomischer  
Expressionsdaten**

Zur Erlangung des akademischen Grades  
Doctor rerum medicarum (Dr. rer. medic.)

vorgelegt der Medizinischen Fakultät der Charité -  
Universitätsmedizin Berlin

von  
Daniel Schober  
aus Ötjendorf, Stormarn

# Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b> .....	<b>12</b>
1.1	Struktur der Arbeit .....	12
1.2	Sprachliche Konventionen.....	12
1.3	Problembeschreibung: Genannotation und Expressionsanalyse.....	13
1.3.1	Heterogenität der Genannotationen .....	14
1.3.2	Mangel an Kontext-Daten.....	15
1.3.3	Primitive Repräsentationsformalismen.....	15
1.3.4	Laborspezifische Annotationen .....	16
1.4	Ziel der Arbeit.....	16
<b>2</b>	<b>Grundlagen</b> .....	<b>19</b>
2.1	Ontologie als Schema zur Wissensrepräsentation .....	19
2.1.1	Ontologiedefinitionen und Anwendungsgebiete.....	19
2.1.2	Ontologie als Kommunikationsstandard und Modell .....	21
2.1.3	Bestandteile ontologischer Formalisierung (KR-Ideome) .....	21
2.1.3.1	Konzepte .....	21
2.1.3.2	Instanzen .....	22
2.1.3.3	Slots, <i>facets</i> und <i>constraints</i> .....	22
2.1.3.4	Frames und Forms.....	24
2.1.4	Objektorientierung und Vererbung .....	24
2.1.5	Datentypen und ihre Repräsentation durch Slot- <i>widgets</i> .....	25
2.1.6	Semantik der Ontologie (OKBC-CLIPS) .....	26
2.1.7	Beschreibung des Wissensbank-Editors Protégé-2000.....	28
2.1.8	Erstellung von Ontologien und <i>ontology engineering</i> -Standards.....	30
2.2	Ontologiebasiertes Wissensmanagement in den Biowissenschaften.....	31
2.3	Anwendungsdomäne: <i>Toll-like Receptors</i> und dendritische Zellen .....	31
2.4	Expressionsanalyse mit dem Affymetrix <sup>®</sup> Human Genome U95Av2 GeneChip <sup>®</sup> .....	33

<b>3</b>	<b>Ergebnisse.....</b>	<b>34</b>
3.1	Erstellung der Gandr-Ontologie .....	34
3.1.1	Anforderungsspezifikation und Kompetenzfragen.....	34
3.1.2	Wissensakquisition .....	35
3.1.3	Manuelle Erstellung einer prototypischen Ontologie.....	36
3.1.4	Erweiterung der Ontologie um domänenspezifisches Vokabular .....	36
3.1.4.1	Erstellung des Ausgangs-Textkorpus zur KR-Ideom-Extraktion.....	36
3.1.4.2	POS-tagging und Extraktion potentieller KR-Ideome .....	37
3.1.5	Taxonomische Integration der Konzepte unter die prototypische Ontologie.....	39
3.1.5.1	Benennungs-Konventionen für Konzepte und Slots.....	40
3.1.5.2	Nutzung einer Text-Konkordanz zur Konzeptpositionierung .....	41
3.1.5.3	Gliederung in <i>top-level</i> -Module .....	42
3.1.5.4	Partonomien und Prozess-Taxonomien .....	43
3.1.5.5	Formalisierung als Konzept, Slot oder Instanz.....	43
3.1.5.6	Nutzung von Mehrfachvererbung.....	44
3.1.6	Erweiterung der Semantik .....	44
3.1.6.1	Hinzufügen von Slots und Relationen .....	44
3.1.7	Integration vorhandener Ontologien.....	45
3.1.8	Entkopplung einer molekularbiologischen <i>upper-level-Ontologie</i> .....	46
3.2	Beschreibung der Gandr-Ontologie.....	47
3.2.1	Grundlegende <i>top-level</i> -Module.....	47
3.2.2	Taxonomische Organisation und Kontext-Einbettung über relationale Slots .....	48
3.2.3	Abbildung ontologischer Ideome auf Protégé-GUI und CLIPS-Format.....	50
3.2.4	Größe und Metrik der Ontologie .....	54
3.2.5	Zugänglichkeit und Veröffentlichungsstatus der Ontologie.....	54
3.3	Erstellung und Beschreibung der Gandr-Wissensbank .....	55
3.3.1	Beschreibung der integrierten Daten (Instanzen) .....	55
3.3.1.1	Verknüpfungen zu externen Daten über Hyperlinks .....	56

3.3.2	Datenimport mit dem Datagenie-Plugin .....	57
3.3.3	Genannotation durch Verschieben ( <i>drag and drop</i> ) von probe set IDs .....	58
3.3.4	Größe der Wissensbank, Systemanforderungen und Performanz.....	59
3.4	Anwendungen der Gandr-Wissensbank.....	60
3.4.1	Formale Annotation von probe set IDs unter Nutzung von Mehrfachvererbung .	61
3.4.2	Integration des Speicher- und Modellierungs-Formates.....	62
3.4.3	Assoziative und kontextsensitive Navigation .....	62
3.4.4	Ontologische Informationsextraktion mit dem Queries & Export-Tab .....	63
3.4.4.1	Export von Anfrageergebnissen.....	66
3.4.5	Visualisierungen der Wissensbank .....	67
3.4.5.1	DAG-basierte Visualisierung mit GraphViz und dem OntoViz-Plugin .....	67
3.4.5.2	Spring-Layout-Visualisierung mit Touch Graph und dem TGViz-Plugin .....	68
3.4.6	Wissensakquisition und Konsistenzprüfung über <i>constraints</i> .....	69
3.4.7	Wissensaustausch und Ontologieexport .....	70
3.4.8	Programmgesteuerte Manipulationen der Wissensbank (JessTab).....	71
3.4.9	PROMPT <i>ontology-versioning</i> und <i>-merging</i> .....	72
3.5	Anpassungen der Benutzeroberfläche (GUI).....	73
3.5.1	Darstellungsoptimierung über Slot- <i>widgets</i> und <i>browser keys</i> im Forms-Tab.....	73
3.6	Modifizieren und Erweitern der Ontologie.....	74
3.6.1	Einführung von Slot-Hierarchien.....	74
3.6.2	Erweiterung der KR-Semantik (Slothierarchien, Metakonzepte und -slots) .....	75
3.7	Die Wissensbank als Internet-Anwendung: WebGandr .....	76
3.8	Aufbau der Gandr-Internetseite, Schulungsfilm und Dokumentation .....	77
<b>4</b>	<b>Diskussion .....</b>	<b>78</b>
4.1	Lexikalische Eigenschaften molekularbiologischer Terminologien.....	78
4.2	Neurokognitive Grundlagen der Wissensakquisition .....	78

4.2.1	Strukturtreue und Kontext erhöhen Interpretationsgeschwindigkeit.....	79
4.2.2	Festigung und Erweiterung des Wissensmodells .....	80
4.3	Anlehnung an <i>ontology engineering</i> -Methodologien (ONIONS) .....	81
4.4	Probleme bei der Erstellung von Ontologien.....	82
4.4.1	Kommunikation mit den Experten und Wissensakquisition .....	82
4.4.2	Taxonomisierungs-Probleme.....	82
4.4.2.1	Konzept oder Instanz, Subkonzept oder Slot.....	83
4.4.2.2	Konzept oder Instanz als Slotwert .....	84
4.4.2.3	Repräsentation von Transformationen und graduellen Zustandsübergängen... ..	84
4.4.2.4	Kontextwandel, Synonyme, Redundanz und taxonomische Inkonsistenzen....	85
4.4.2.5	Gleiche Detailliertheit bei Geschwisterkonzepten.....	86
4.4.3	Fehler in zu integrierenden Ontologien .....	86
4.5	Beurteilung der Gandr-Ontologie .....	87
4.5.1	Ontologie-Typ.....	87
4.5.2	Beurteilung der Kodierungssprache und Expressivität.....	89
4.5.2.1	RDB vs. CLIPS vs. OWL .....	90
4.6	Beurteilung der Gandr-Wissensbank.....	90
4.6.1	Formale Klassifizierung der Anwendung nach dem System Uscholds.....	91
4.6.2	Beurteilung der Anwendung anhand der Anforderungsspezifikation .....	94
4.6.3	Beurteilung der IR-Kapazität.....	94
4.6.4	Dokumentation und Schulung .....	95
4.6.5	Akzeptanz beim Nutzer .....	96
4.7	Beurteilung der Visualisierungsansätze.....	97
4.7.1	Datengetriebene und konfigurierbare GUI .....	97
4.7.2	Visualisierungen der Wissensbank-Inhalte.....	98
4.7.2.1	Vorteile der Frames gegenüber tabellarischen Darstellungen .....	98
4.7.2.2	Vergleich mit Kohns <i>molecular interaction maps</i> .....	99

4.8	Vergleich mit anderen Annotations-Systemen und Ontologien .....	100
4.8.1	Affymetrix®-eigene Annotationsmöglichkeiten .....	101
4.8.2	Gene Ontology, GONG und GO-Mining-Tool.....	101
4.8.3	UMLS .....	103
4.8.4	MGED, MIAME und MAGE .....	104
4.8.5	TAMBIS .....	105
4.9	Ausblick .....	106
4.9.1	Ontologie-induzierte Konzept-Ikonographien .....	106
4.9.2	Internetseiten-Annotation im <i>semantic web</i> -Ansatz .....	107
4.9.3	Diskriminanzanalysen, Gruppierungsverfahren und maschinelles Lernen .....	108
4.10	Zusammenfassung.....	110
	Literaturverzeichnis.....	112
	Abkürzungen.....	122
	Abbildungsverzeichnis.....	124
	Danksagungen.....	125
	Curriculum Vitae.....	126
	Publikationsliste.....	128
	Erklärung.....	129
	Anhang.....	130