Medical Infobahn for Europe
A. Hasman et al. (Eds.)
IOS Press, 2000

MUSTANG: Wiederverwendbare UMLS-basierte terminologische Dienste

[1] Josef INGENERF, [2] Jörg REINER
[1] *Institut für Medizinische Informatik, Universität zu Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany (ingenerf@medinf.mu-luebeck.de)*
[2] *GSF-Medis-Institut, Postfach 1129, 85758 Neuherberg, Germany (reiner@gsf.de)*

Abstract. Das UMLS-System (Unified Medical Language System) als weltweit umfangreichste Terminologie-Ressource wird bis heute in Deutschland weniger als erwartet eingesetzt. Sowohl ein effektiver Zugriff auf Online-Wissen mit den Standard-Problemen des Information Retrieval (Vollständigkeit, Korrektheit) als auch die Integration heterogener Anwendungssysteme können von UMLS-basierten Dolmetscher-Diensten profitieren. Das MUSTANG-System (Medical UMLS based Terminology Server for Authoring, Navigating and Guiding the Retrieval to Heterogeneous Knowledge Sources) bietet genau solche Dienste an. Auf mehreren technischen Ebenen (z.B. Web-basierter Zugang, CORBA-basierte Schnittstelle) lassen sich diese Dienste abgreifen. Die sogenannten "Lexical Services" des CORBAmed-Standards wurden bei der Spezifikation der Dienste berücksichtigt. Die generelle Bereitschaft, autonom entwickelte Systeme zu integrieren beziehungsweise deren Dienste wiederzuverwenden hängt unter anderem davon ab, ob die jeweilige Autonomie zugunsten von Standards zur Wahrung einer übergreifenden Konsistenz und Transparenz aufgegeben wird. Über erste Anwendungen des MUSTANG-Systems wird berichtet.

1. Einleitung

Im Rahmen des vom BMBF geförderten MEDWIS-Projekt „Medizinische Wissensbasen" unterstützt der Arbeitskreis "Terminologie" die an der Entwicklung von wissensbasierten Systemen arbeitenden Einzelvorhaben. Medizinische Terminologie-Standards spielen eine elementare Rolle für die referentielle Integrität zwischen Wissensbasen und Patientendaten, für die Ergänzung der Funktionalität wissensbasierter Systeme um effektive Zugriffe auf Online-Wissen sowie für die Integration von wissensbasierten Systemen in klinische Anwendungssysteme. Hierüber wurde bereits berichtet [1]. Nicht jedes Projekt muß eine Komponente zur terminologischen Standardisierung immer wieder neu entwickeln. Es wurde erkannt, daß das UMLS-System (Unified Medical Language System) eine ausgezeichnete Grundlage für das Angebot terminologischer Dienste ist. Dieses bietet die Integration nahezu aller weltweit relevanten medizinischen Terminologiesysteme wie die ICD-Klassifikation, den MeSH-Thesaurus, usw. Mit dem MUSTANG-System (Medical UMLS based Terminology Server for Authoring, Navigating and Guiding the Retrieval to Heterogeneous Knowledge Sources) werden die recht umfangreichen und komplexen UMLS-Ressourcen effizient verwaltet und über standardisierte Schnittstellen dem Interessenten angeboten.

2. Terminologische Dienste

Über die Integration von wissensbasierten Systemen hinaus ist ein Terminologieserver in einem weiteren Kontext zu sehen; nämlich als integraler Bestandteil einer Telemedizin-Infrastruktur, wie sie z.B. in der Roland Berger Studie [2] beschrieben wurde. Die Notwendigkeit eines Terminologieservers zur Lösung des zunehmenden semantischen Interoperabilitätsproblems zwischen medizinischen Anwendungssystemen ist unbestritten. Die folgende Abbildung 1 wurde der Studie entnommen und erweitert. An den Ecken wurden Szenarien positioniert, die sich von a) nach d) jeweils hinsichtlich ihrer Komplexität unterscheiden.

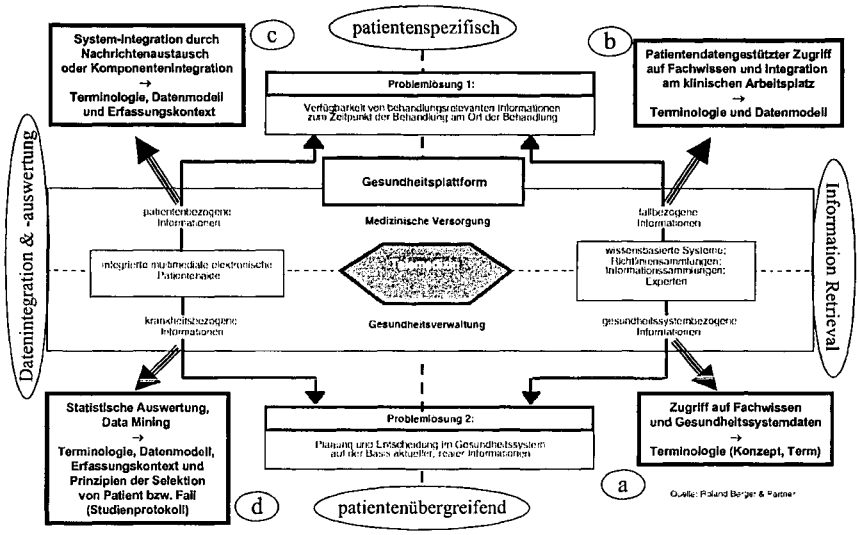

Abb.1 Varianten semantischer Interoperabilität

Zunächst läßt sich festhalten, daß verfügbares terminologisches Wissen den Zugriff auf Fachwissen und Gesundheitssystemdaten verbessern kann (a). Ein MeSH-Thesaurus etwa ermöglicht verbesserte Informationsrecherchen in der MEDLINE-Literaturdatenbank oder in WEB-basierten Informationsangeboten [3]. Eine Integration eines solchen Information Retrieval Systems in klinische Anwendungssysteme erlaubt einen patientendaten-gestützten Zugriff auf Fachwissen (b). Dieses verlangt eine zusätzliche Kenntnis des Datenmodells. Zum Beispiel kann ausgehend von einer ICD-Kodierung ein Zugriff auf MEDLINE erfolgen, da im Hintergrund auf einen korrespondierenden MeSH-Kode abgebildet wird. Weitere Datenelemente (z.B. Diagnosenzusätze wie "Ausschluß von") müssen bei einer solchen integrativen Vorgehensweise jedoch berücksichtigt werden.

Eine ganz andere Komplexitätsstufe ergibt sich, wenn es um die Kommunikation, Integration und rechnergestützte Weiterverarbeitung von Patientendaten geht (c). Sowohl für eine "loose Kopplung" zweier Systeme mittels Nachrichtenaustausch als auch für eine "enge Kopplung" mittels Systemintegration muß eine systemübergreifende Datenkonsistenz gewährleistet werden [4]. Hierzu müssen die Datenmodelle zweier i.a. autonom entwickelter Systeme mit ihren jeweiligen Anwendungslogiken aufeinander bezogen werden. Eine zuverlässige Interpretation und Verarbeitung von Patientendaten eines anderen Systems muß den Erfassungskontext berücksichtigen. Das Szenario (c) in Abb. 1 ist dadurch gekennzeichnet, daß neben der syntaktischen und semantischen Bearbeitungsebene eine pragmatische Bearbeitungsebene ergänzt werden muß. Eine zuverlässige Integration von Patientendaten aus einem in ein anderes Anwendungssystem mit unterschiedlichen Datenbankschemata ist nur bis zu einem gewissen Grad automatisierbar [5]. Am Beispiel der konkreten Integration eines wissensbasierten Systems in ein klinisches Anwendungssystem zeigen z.B. Eich et al. [6] recht deutlich, daß die Abbildung zwischen Wertebereichen sehr viel "Handarbeit" erfordert. Der Vollständigkeit halber verlangt das letzte Szenario in Abb.1 neben einer festgelegten Terminologie, Datenmodell sowie dem Erfassungskontext auch eine Festlegung der Patienten- bzw. Fallselektion (d). Genau dieses ist Gegenstand von Studienprotokollen.

3. Terminologieserver

Terminologieserver realisieren einfache und komplexe Dienste auf der Basis von Wissen um Begriffe, um Bezeichnungen und Kodes. Konkurrierende Ansätze von Terminologieservern unterscheiden sich im wesentlichen durch unterschiedliche Terminologie-Ressourcen.

3.1 UMLS

Die National Library of Medicine (NLM) in den USA initiierte Ende der Achtziger Jahre das Unified Medical Language System (UMLS) - Projekt mit genau dem Ziel, das semantische Interoperabilitätsproblem in verteilten, heterogenen Anwendungssystemen in der Medizin zu lösen [7]. In einem Metathesaurus werden aus ca. 50 der weltweit wichtigsten Kodesysteme (auch Source-Vocabularies genannt) alle enthaltenen Kodes auf eindeutig identifizierte Konzepte abgebildet, die über eine CUI (Concept Unique Identifier) Kodesystem-unabhängig identifiziert werden. UMLS "erfindet" also keine eigenen Kodes, sondern bildet einen Mehrwert durch die Zusammenführung aller Kodes der beteiligten Kodesysteme mit ihren verschiedenen (u.a. multilingualen) Bezeichnungen. Gleichzeit werden die Konzepte durch ein seperates semantisches Netz getypt. Die Abb.2 stellt einige wichtige Informationen zusammen, die zu einem Begriff wie "Diabetische Nephropathie" aus dem UMLS angeboten werden können. Daraus wird bereits umgekehrt deutlich, welches die elementaren terminologischen Dienste sind, die ein UMLS-basierter Terminologieserver anbieten kann. Eine Kode-Konversion wird ermöglicht durch die Tatsache, daß zu einem ICD9-Kode "250.4" das zugeordnete Konzept "CUI C0011881" aufgesucht wird, dem wiederum der MeSH-Kode "C12.777.419.192" zugeordnet ist. Aufwendige Algorithmen sind jedoch notwendig, weil eben nicht zu jedem Kode eines Kodesystems ein begrifflich identischer Kode eines anderen Kodesystems existiert. Für diese und andere Aufgabenstellungen können die im UMLS-System bereitgestellten Begriffsrelationen genutzt werden (siehe Abb.3).

Definition (MeSH 98):

Includes renal arteriosclerosis, renal arteriolosclerosis, Kimmelstiel-Wilson syndrome (intercapillary glomerulosclerosis), acute and chronic pyelonephritis, and kidney papillary necrosis in individuals with diabetes mellitus.

Syntaktischer Typ:	Semantischer Typ:
nominal phrase	Disease or Syndrome

Bezeichnungen bzw. Synonyme:		Kodierungen:	
Englisch:	Diabetes, nephropathy	MeSH-Thesaurus:	C12.777.419.192
	Diabetes with renal manifestations		C18.452.297.402
German:	Diabetische Nephropathie	ICD9-Klassifikation:	250.4
Italienisch:	Nefropatie diabetiche	Read Code:	X30Kk
Spanisch:	Nefropatias Diabeticas	SNOMED International:	DB-62100
Russian:	Diabeticheskie Nefropatii	Crisp Thesaurus:	0862-6260
...		....	

Abb.2 Begriffsattribute zum Konzept "Diabetische Nephropathie (CUI C0011881)"

Die meisten Begriffsrelationen in UMLS entstammen den Hierarchien der jeweiligen Kodesysteme. Es handelt sich um Ober-/Unterbegriffsbeziehungen, die nicht näher in generative ("isa") und partitative ("part_of") Beziehungen differenziert werden (A). Weiterhin existiert ein semantisches Netz, welches z.B. Typen wie "Disease or Syndrome" über eine Relation "located_in" mit dem Typ "Body Location" verknüpft (B). Drittens enthält das UMLS über eine Auswertung der MeSH-Kodierung in der MEDLINE-Literaturdatenbank

eine statistische Aussage über das gemeinsame Vorkommen zweier MeSH-Kodes bzw. der korrespondierenden Konzepte, die sogenannten Co-Occurences (C).

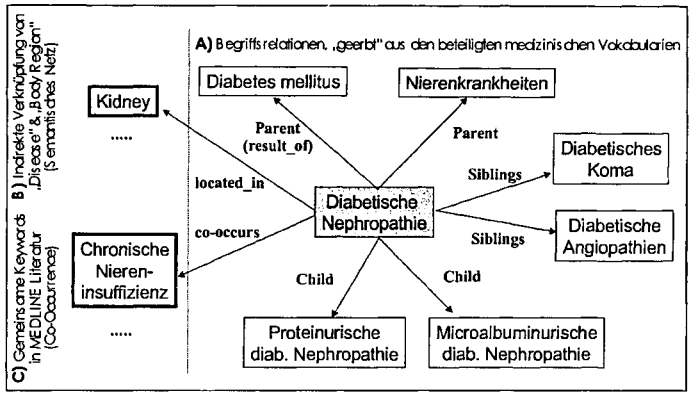

Abb.3 Begriffsrelationen zum Konzept "Diabetische Nephropatie (CUI C0011881)"

3.2 MUSTANG

Die jährlichen Aktualisierungen der UMLS-Terminologie-Ressourcen werden in ein ORACLE-Datenbanksystem eingespielt. Aufgrund des enormen Umfanges (physikalisch handelt es sich um ca. 4 GB "Rohdaten") ist ein geeignetes Datenbankschema mit entsprechender Indizierung eine Voraussetzung für das weitere Vorgehen. Die Architektur des MUSTANG-Systems ist bestimmt durch die Repräsentation des UMLS-Datenbestandes sowie möglicher Ergänzungen, durch die Bereitstellung von terminologischen Diensten via einer standardisierten Schnittstelle sowie einer effizienten Implementierung dieser Dienste.

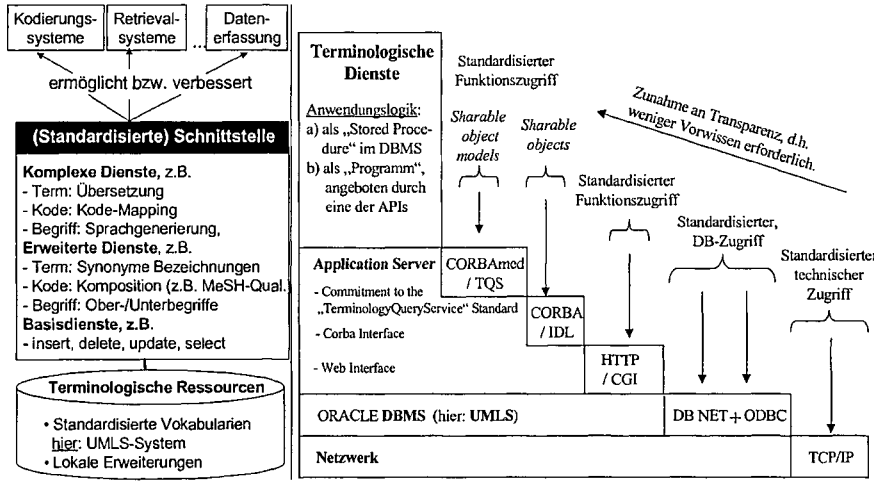

Abb.4 Client-Server Architektur des MUSTANG-Systems

Das MUSTANG-System stellt Daten und Funktionen über mehrere Schnittstellen zur Verfügung. Einerseits wird eine Web-basierte Schnittstelle angeboten, mit der man interaktiv die zahlreichen Informationen recherchieren und navigieren kann. Hier gibt es

Spezialfunktionen wie die Möglichkeit des Downloads von Suchergebnissen oder die Nutzung der Ergebnisse als Eingabe in Web-basierten Suchmaschinen. Für eine System-Integration terminologischer Dienste in Fremdanwendungen wurde eine CORBA-Schnittstelle implementiert. Diese CORBA-Schnittstelle orientiert sich an das Objektmodell der "Terminology Query Services" der CORBmed-Initiative [8].

4. Anwendungen und Erfahrungen mit dem MUSTANG-System

Es gibt ganz grundsätzlich ein Beharrungsvermögen, verteilte Dienste wiederzuverwenden und in autonome Systeme zu integrieren. Die wichtigsten Gründe betreffen die Qualität der bereitgestellen Funktionen, softwaretechnische Schnittstellenfragen, Fragen der Verfübarkeit und Effizienz bis hin zur grundsätzlichen Bereitschaft, die eigene Autonomie aufzugeben zugunsten der Einbindung standardisierter Funktionen [9]. Für kommerzielle Interessenten am MUSTANG-System sind schließlich auch lizenzrechtliche Fragen sehr wichtig, inkl. der Verwendbarkeit der in UMLS enthaltenen Vokabularien. Bespielhaft seien folgende Nutzungen des MUSTANG-Systems genannt:

- Unter der Adresse http://mustang.gsf.de wird ein interaktives Zugangssystem zum UMLS-System angeboten.
- Über die reine Recherche und Navigation in den UMLS-Ressourcen hinaus lassen sich unter dem Menü "Wissensressourcen" die Rechercheergebnisse über "Add to Query" für den Aufbau eines Query-Strings mit der entsprechenden Suchmaschinen-Syntax für Zielsysteme wie Medline, Alta Vista, Medivista usw. nutzen (siehe Abb.1, Szenario a).
- Im Rahmen des Projektes TEDI – Telematik für Ernährung und Diätetik – ist das wissensbasierte Informationssystem MIEL (Multimediales Informationssystem für Ernährungsfragen und Lebensmittelauswahl) entwickelt worden. MIEL stützt sich intern auf eine terminologischen Komponente, die mit Hilfe von MUSTANG erarbeitet wurde.
- Auch im MEDWIS-Projekt "ESAB" hat sich eine Terminologiekomponente zur Abstimmung von Dokumentationssystem und Wissenssystem bewährt. Aus Gründen der Wartung und Fortschreibung der ESAB-Terminologie-Komponente soll dieser Teil durch CORBA-basierte MUSTANG-Dienste ersetzt werden. Ein erster Test auf der technischen Ebene konnte erfolgreich abgeschlossen werden.
- Ein System zur Definition und Überwachung von HL7-basierten klinischen Systemkopplungen stützt sich zentral auf die Abbildung zwischen je zwei Datentypen mit ihren Wertebereichen. Die jeweiligen systemeigenen Stammtabellen einerseits und die standardisierten "HL7-Coding Tables" können mit MUSTANG-Diensten aufeinander bezogen werden. Die HL7-Terminologie ist in UMLS enthalten und wird lokal erweitert.

Danksagung

Dem Bundesministerium für Forschung und Technologie wird für die Unterstützung im Rahmen des MEDWIS-Programmes (hier: des Arbeitskreises „Terminologie“) gedankt.

Literatur

[1] Ingenerf J. Interoperabilität zwischen medizinischen Anwendungssystemen. Informatik, Biometrie und Epidemiologie in Medizin und Biologie 1998; *29* (1): 69-76.

[2] RolandBerger&PartnerGmbH. *Telematik im Gesundheitswesen - Perspektiven der Telemedizin in Deutschland.* 1998.

[3] Joubert M, Fieschi M, Robert J-J, *et al.* UMLS-based Conceptual Queries to Biomedical Information Databases: An Overview of the Project ARIANE. J. Am. Med. Inform. Assoc. 1998; 5: 52-61.

[4] Lenz R, Blaser R, Kuhn KA. Hospital Information Systems: Chances and Obstacles on the Way to Integration. In: Kokol P, al. e, eds. *Proc.MIE'99.* Amsterdam: IOS-Press, 1999: 25-30.

[5] Ingenerf J. Telemedicine and Terminology: Different Needs of Context Information. IEEE Transactions on Information Technology in Biomedicine 1999; *3* (2): 92-100.

[6] Eich H-P, Lang K, Ganslandt T, *et al.* Meta Data Dictionary to Link and Reuse Knowledge-based Systems in Medicine. In: Puppe F, al. e, eds. *Proc. of the XPS '99.* Berlin: Springer, 1999: 10 p.

[7] Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Methods of Information in Medicine 1993; *32*: 281-291.

[8] OMG (Object Management Group). *CORBAmed, WG "Lexicon Services" (Interface Definition language (IDL) for Interoperability in healthcare vocabulary systems*: 1997.

[9] Goodhue DL, Wybo MD, Kirsch LJ. The Impact of Data Integration on the Costs and Benefits of Information Systems. MIS Quarterly (http://www.misq.org) 1992; *16* (3): 293-322.