Onto-Builder: Software-Werkzeug für den Aufbau von Data Dictionaries

B. Heller, K. Kühn, M. Löffler

Universität Leipzig
Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE)
Härtelstraße 16-18, 04107 Leipzig
{barbara.heller, katrin.kuehn, markus.loeffler}@imise.uni-leipzig.de

Zusammenfassung. In vielen Fachgebieten sind präzise Begriffsdefinitionen äußerst wichtig, um eine eindeutige Kommunikation und Interoperabilität zu ermöglichen. Vor diesem Hintergrund entstand die Idee des *Onto-Builders*, dessen Ziel zunächst darin besteht, die Definitionsalternativen von Begriffen zu sammeln und als Data Dictionary zur Verfügung zu stellen. Zusätzlich soll die Analyse der Definitionen in Bezug auf ihre Gemeinsamkeiten und Unterschiede sowie deren Harmonisierung unterstützt werden. Standardisierte Begriffsdefinitionen werden jedoch nicht erzwungen, da die Unterschiede in Definitionen inhaltlich gerechtfertigt sein können, z.B. aufgrund der Verwendung in unterschiedlichen Fachgebieten, durch studienspezifische Bedingungen oder verschiedene Expertensichten.

1 Problemstellung

Klinische Studien werden zunehmend multizentrisch und auf internationaler Ebene durchgeführt. Damit steigen die Herausforderungen bezüglich der Umsetzung geltender Qualitätsstandards, die in nationalen sowie internationalen Richtlinien und Gesetzen verankert sind.

Präzise Begriffsdefinitionen sind in diesem Umfeld von großer Bedeutung, da sie für eine objektive Datenerfassung und -auswertung sowie eine eindeutige Kommunikation zwischen den teilnehmenden Einrichtungen unerlässlich sind. Zudem ermöglichen sie externen Experten, die Forschungsergebnisse einer Studie korrekt zu interpretieren und anzuwenden.

Allerdings weisen viele klinische Studien Defizite in diesem Punkt auf. Die Hauptprobleme und deren Konsequenzen werden im Folgenden erläutert.

1.1 Fehlende oder ungenaue Begriffsdefinitionen

Begriffe werden oft ungenau definiert (z.B. in Studienprotokollen), oder sie werden sogar verwendet ohne sie explizit zu definieren. Ein Beispiel für ungenaue Begriffsdefinitionen sind fehlende präzise Angaben zur Durchführung von Untersuchungen (z.B.

welches Messverfahren oder welche Untersuchungsmethode anzuwenden ist). Ein häufiges Beispiel ist auch die unzureichende Definition von Intervallen, indem keine Aussage über die Zugehörigkeit der Intervallgrenzen getroffen wird.

Fehlende oder ungenaue Begriffsdefinitionen führen zu einer schlechten Nachvollziehbarkeit von Studienergebnissen. Des Weiteren werden mehrdeutige Interpretationen eines Begriffs ermöglicht, die z.B. die notwendige Beobachtungsgleichheit der Patienten in einer Studie gefährden.

1.2 Mehrfache Verwendung von Begriffen und Definitionen

Häufig kommt es vor, dass der gleiche Begriff innerhalb einer Studie mehrfach verwendet wird, z.B. im Studienprotokoll, in Dokumentationsbögen, Standardarbeitsanweisungen, Software-Anwendungen, usw. Dabei können Inkonsistenzen auftreten, z.B. wenn die zugrunde liegenden Definitionstexte unabhängig voneinander bearbeitet werden. In diesem Zusammenhang können Software-Werkzeuge helfen, den manuellen Aufwand zu verringern und die dokumentenübergreifende Konsistenz sicherzustellen.

1.3 Uneinheitliche Definitionen

Begriffe werden in verschiedenen Studien oft unterschiedlich definiert, obwohl standardisierte Definitionen im Hinblick auf einen weitreichenden Austausch von Daten und Ergebnissen wünschenswert sind. Als Beispiele sind hier unterschiedliche Benennungen für gleiche Begriffe zu nennen, sowie die Verwendung unterschiedlicher Messverfahren, Einheiten oder Datenformate.

Inhaltliche Unterschiede in den Definitionen treten aber ebenfalls häufig auf und haben zur Folge, dass Ergebnisse aus verschiedenen Studien oft nicht miteinander vergleichbar sind. Diese Problematik wird auch in [1] analysiert. Im Rahmen eines internationalen Workshops wurden dort Definitionen für Kriterien zur Beurteilung des Therapieerfolges (z.B. „Komplette Remission“, „Partielle Remission“) in Studien zu Non-Hodgkin-Lymphomen (NHL; bösartige Tumore des lymphatischen Systems) verglichen. Dabei wurde festgestellt, dass jede größere Studiengruppe eigene Definitionen benutzt. So unterscheiden sich z.B. die Definitionen für „Komplette Remission“ in der geforderten Dauer der Symptomfreiheit oder in den geforderten Untersuchungen zur Feststellung einer kompletten Remission. Das interessante Fazit dieser Arbeit besteht darin, dass die aktuell verfügbaren wissenschaftlichen Erkenntnisse nicht ausreichen, um eine bestimmte Definition als „korrekt“ zu bestätigen. Stattdessen ist höchstens eine Einigung auf die Verwendung einer Definitionsvariante möglich, wobei selbst dies aufgrund widersprüchlicher Expertenmeinungen oft schwierig oder unmöglich ist.

1.4 Notwendigkeit verschiedener Sichten

Obwohl eine Vereinheitlichung von Begriffsdefinitionen aus den genannten Gründen generell wünschenswert ist, muss bedacht werden, dass dies nicht immer sinnvoll ist. Je nach Kontext können Definitionen verschiedener Detailliertheit angemessen sein, z.B. kann der Begriff „Remission“ allgemeingültig für den Kontext Medizin als „Zurückgehen von Krankheitserscheinungen“ ([2], S. 1365) definiert werden, im Kontext

einer konkreten Studie ist jedoch eine sehr präzise, krankheitsspezifische Definition nötig.

Überdies sind viele Fachgebiete interdisziplinär, d.h. Experten verschiedener Fachgebiete arbeiten zusammen (z.B. Radiologen, Pathologen, Dokumentare, Biometriker im Rahmen einer klinischen Studie) und benötigen u.U. unterschiedliche Informationen zu einem Begriff, die sich in unterschiedlichen Definitionsvarianten niederschlagen können. Beispielsweise ist ein Biometriker vor allem an Informationen bezüglich der statistischen Auswertung interessiert, die für einen Pathologen irrelevant sein können.

2 Stand der Forschung

Der Bedarf einheitlicher Begriffsdefinitionen für die Unterstützung einer eindeutigen Kommunikation ist in vielen Wissensgebieten anerkannt. Insbesondere in der Medizin gibt es schon seit langem verschiedene terminologische Standardisierungsbemühungen (z.B. ICD, SNOMED, LOINC, GALEN, etc.). Bisherige Ansätze zur Terminologiestandardisierung haben jedoch entscheidende Nachteile im Bezug auf die beschriebene Problematik.

Zum einen mangelt es im Rahmen der Erstellung von Terminologien an Softwareunterstützung für den Prozess der Standardisierung und Konsensfindung. Zum anderen ist die Genauigkeit der Begriffsdefinitionen in vorhandenen Terminologien nicht ausreichend oder Definitionen fehlen völlig, wodurch wiederum die gewünschte eindeutige Kommunikation behindert wird.

Ein weiterer Aspekt ist die nicht ausreichende ontologische Fundierung, die selbst bei etablierten Begriffsystemen in der Medizin festzustellen ist [3]. Meist fehlen semantisch definierte Relationen zwischen Begriffen sowie formale Begriffsdefinitionen. Dies sind jedoch wesentliche Voraussetzungen für eine semantisch korrekte Wissens- bzw. Informationsverarbeitung und -wiederverwendung.

Ferner wird die Machbarkeit einer Standardisierung vorausgesetzt und daher keine Möglichkeit geboten, voneinander abweichende alternative Definitionen eines Begriffes zu repräsentieren. Dies ist jedoch, wie im vorigen Abschnitt erläutert, notwendig, da ein Konsens nicht immer möglich ist.

Schließlich wurden die meisten Terminologien für einen bestimmten Anwendungszweck entwickelt. Dies manifestiert sich z.B. in starren Begriffshierarchien, in denen Begriffe bezüglich einer bestimmten Sicht klassifiziert werden oder sogar verschiedene Sichten in einer Hierarchie vermischt werden. Daraus kann inkorrekte Vererbung resultieren, und die Wartbarkeit und Wiederverwendbarkeit wird deutlich erschwert (vgl. [4]).

3 Zielsetzung

Vor diesem Hintergrund entstand die Idee, das Software-Werkzeug *Onto-Builder* zu entwickeln, um den Aufbau und das Management von Data Dictionaries zu unterstützen. Unter einem Data Dictionary verstehen wir eine zentrale und wiederverwendbare Begriffsbasis, die in mehreren Einrichtungen in einem bestimmten Fachgebiet einge-

setzt werden kann und eine einheitliche, präzise Begriffsgrundlage bildet (z. B. für verschiedene Dokumente, Softwareanwendungen, etc.).

Allerdings wird eine Standardisierung nicht erzwungen. Stattdessen soll der Prozess der Standardisierung und der dazu notwendigen Konsensfindung unterstützt werden, aber auch eine Möglichkeit für die Repräsentation kontextabhängiger Begriffsdefinitionen und verschiedener Expertensichten geboten werden, falls eine Vereinheitlichung aus fachlichen Gründen nicht möglich oder nicht sinnvoll ist.

Ein wichtiges Ziel bei der Entwicklung des *Onto-Builders* ist die Wahrung der Domänenunabhängigkeit der Software. Das Tool soll nicht auf die Domäne der klinischen Studien beschränkt werden, sondern problemlos auch in anderen Domänen einsetzbar sein, da die Problematik von präzisen Begriffsdefinitionen in nahezu allen Fachgebieten von Bedeutung ist und zunehmend an Aufmerksamkeit gewinnt.

4 Das Software-Werkzeug *Onto-Builder*

4.1 Umsetzung

In einer ersten Phase wurde der *Onto-Builder* als Internetapplikation implementiert [5] und unterstützt die Sammlung von Begriffsdefinitionen, die vergleichende Analyse der Unterschiede, sowie die Konsensfindung. Abbildung 1 zeigt einen Screenshot der Anwendung. Es ist der Term „Progress“ zu sehen, zu dem drei Definitionsvarianten eingegeben wurden, die sich in ihrer Genauigkeit und vor allem in ihrer Domänenspezifität stark unterscheiden (vgl. Abschnitt 1.4).

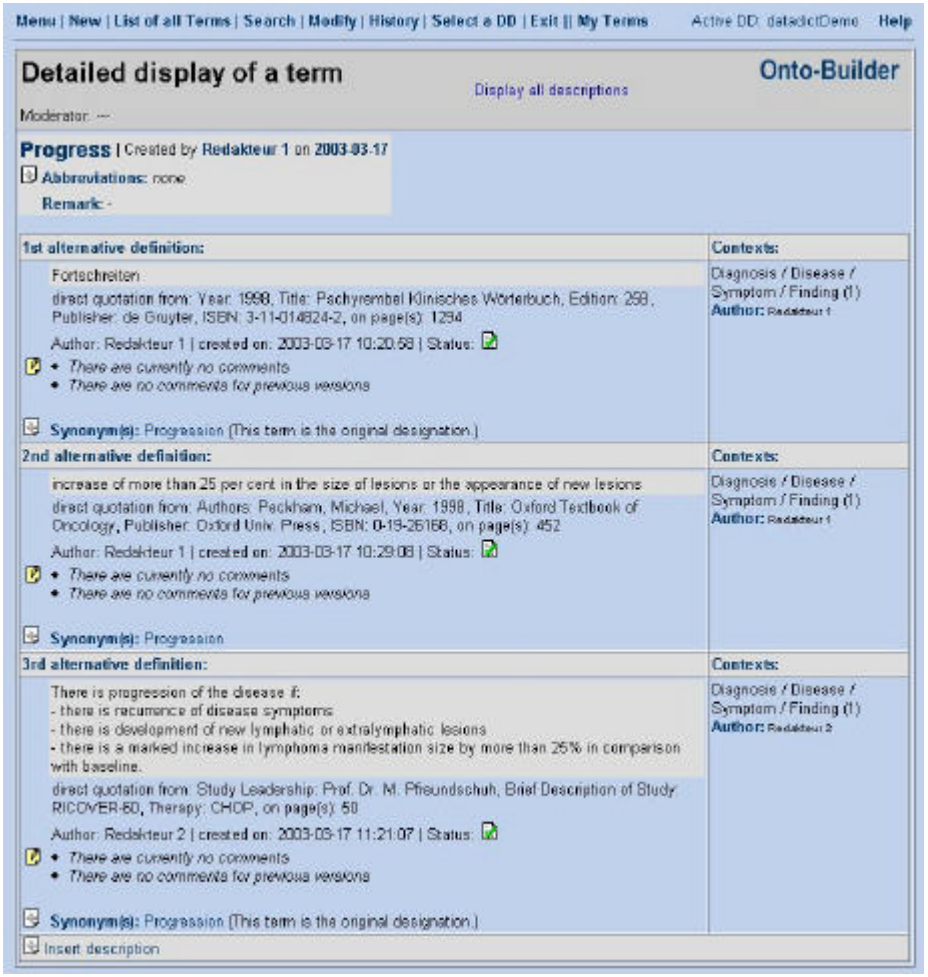
Folgende Inhalte können in dieser ersten Version des *Onto-Builders* erfasst werden:

- Terme (d.h. Benennungen von Begriffen) und deren mögliche Abkürzungen
- Sprache eines Terms
- Definitionsvorschläge (entweder eigene Definitionen oder Definitionen aus Literaturquellen, Studienprotokollen, etc.)
- Quellen der Definitionen zwecks Rückverfolgbarkeit
- Kontexte, in denen eine Beschreibung zu einem Begriff gültig ist
- Synonyme (in Abhängigkeit der Definitionsalternativen)
- Konsensvorschläge für Definitionen von Begriffen
- Autoreninformationen zu den Inhalten

Neben den Basisfunktionen wie Anlegen und Ändern der o.g. Inhalte weist der *Onto-Builder* folgende Funktionen auf:

- Kommentierung eingegebener Beschreibungen
- Versionierung von Beschreibungen, um Änderungen nachvollziehen zu können
- Zugriffsschutz und Nutzergruppenverwaltung (Leser, Redakteur, Moderator, Administrator), d.h. einige Funktionen stehen nur für bestimmte Nutzergruppen zur Verfügung
- Unterstützung von anwenderspezifischen Zugriffsmodalitäten (z.B.: eine Definition darf nur von deren Autor geändert werden)

Abb. 1 Screenshot des *Onto-Builders*: Drei Definitionsvarianten des Terms „Progress".



Der *Onto-Builder* wird von verschiedenen Anwenderkreisen genutzt, darunter das Koordinierungszentrum für Klinische Studien Leipzig (www.kksl.uni-leipzig.de) sowie u.a. die Kompetenznetze Maligne Lymphome (www.lymphome.de) und Herzinsuffizienz (www.knhi.de). Die dabei gewonnen Erfahrungen sowie die Analyse der Begriffsdefinitionen fließen in die Weiterentwicklung des *Onto-Builders* ein.

4.2 Weiterentwicklung

Aufbauend auf einer Evaluation der ersten Version sowie der Analyse eingegebener Definitionen wurde das Datenmodell für die Repräsentation der Data Dictionary Inhalte deutlich erweitert. Es bietet Verbesserungen im Hinblick auf die Abbildung und

Strukturierung von Begriffsdefinitionen, z.B. durch die Möglichkeit, für bestimmte Gruppen von Begriffen (z.B. Laborparameter) Definitionsschablonen zu definieren. Dies trägt zu einer Vereinheitlichung der zu vergleichenden Definitionen bei und erleichtert somit die Konsensfindung. Des Weiteren werden zusätzliche Funktionalitäten angeboten wie z.B. die Abbildung von Relationen, Kontexten und Sichten sowie ein detaillierter Qualitätssicherungszyklus bei der Eingabe von Inhalten. Dieses erweiterte Datenmodell wird derzeit implementiert.

Zusätzlich wird im Rahmen der Forschungsgruppe Onto-Med (Ontologies in Medicine [6]) an einer ontologischen Fundierung der Begriffsdefinitionen gearbeitet. Die seit 1998 bestehende Forschungsgruppe ist eine Kooperation des IMISE (Institut für Medizinische Informatik, Statistik und Epidemiologie, Universität Leipzig) und der Abteilung Formale Konzepte des Instituts für Informatik, Universität Leipzig und beschäftigt sich mit der Theorie-Entwicklung im Bereich Ontologie sowie deren Implementierung und praktischer Anwendung in Informationssystemen, insbesondere auf dem Gebiet der Medizin. Im Rahmen der Theorie wird an einem System von Top-Level Ontologien gearbeitet, die domänenunabhängige Begriffe (z.B. Raum, Zeit, Prozess, Kategorie) definieren und diese in einem formalen Modell zur Verfügung stellen [7]. Die Entwicklung von Domänenontologien wird unterstützt, indem einerseits die entwickelten Kategorien und die für sie definierten Zusammenhänge in Form von Axiomen als Grundlage für die Definition domänen-spezifischer Begriffe dienen, während andererseits auf methodologische Ansätze und Vorgehensmodelle zurückgegriffen werden kann [8]. Ziel ist dementsprechend, die in der dabei entwickelten Top-Level Ontologie definierten generischen Begriffe als Definitionsgrundlage für die domänenspezifischen Begriffe im Data Dictionary einzusetzen.

Geplant ist des Weiteren eine Integration des Data Dictionary in die ebenfalls in der Forschungsgruppe Onto-Med entwickelten Softwaretools SOP-Creator [9] und OncoWorkstation [10]. Der SOP-Creator ist ein web-basiertes Werkzeug zur verteilten, qualitätsgesicherten Erstellung und Präsentation von Standardarbeitsanweisungen (engl.: standard operating procedure, SOP), mit einer zentralen Verwaltung der SOPs. Die ebenfalls web-basierte Anwendung OncoWorkstation unterstützt die Auswahl, Planung und Ausführung von studienprotokollbasierten, onkologischen Behandlungen. Beide Anwendungen basieren zum Teil auf den gleichen Begriffen, die in einem zugrunde liegenden Studienprotokoll definiert sind. Eine Schnittstelle, die die Verwendung von Begriffen und deren Definitionen im Data Dictionary in anderen Anwendungen ermöglicht soll die Wiederverwendung der Begriffsdefinitionen erleichtern und die semantische Konsistenz sicherstellen.

5 Diskussion

Auch in dem weiterentwickelten Datenmodell sind einige konzeptuelle Fragestellungen noch ungelöst. Dazu gehört die zu untersuchende Problematik, wie Vererbungshierarchien in Verbindung mit alternativen Begriffsdefinitionen umgesetzt werden können. Üblicherweise werden Begriffe in Vererbungshierarchien angeordnet, wobei Attribute von übergeordneten Begriffen an untergeordnete Begriffe vererbt werden. Da Attribu-

te jedoch zur Definition eines Begriffs beitragen, ist unklar wie eine Vererbungshierarchie in Verbindung mit alternativen Begriffsdefinitionen umgesetzt werden kann.

Die Vorteile der Verfügbarkeit eines Data Dictionary sind dennoch bereits im derzeitigen Stadium ersichtlich. Die expliziten und präzisen Begriffsdefinitionen bilden einen Beitrag zur Qualitätssicherung und -verbesserung. Die Beobachtungsgleichheit in klinischen Studien wird sichergestellt, da Interpretationsspielräume minimiert werden. Des Weiteren werden studienübergreifende Auswertungen ermöglicht, da die einer Studie zugrunde liegenden Begrifflichkeiten klar definiert sind.

Auf organisatorischer Seite könnte der Prozess des Aufbaus von Data Dictionaries allerdings noch besser unterstützt werden. Hier werden Möglichkeiten des Einsatzes automatischer Methoden zur Extraktion von Begriffen und Definitionen aus vorhandenen Texten analysiert. Dadurch soll der manuelle Aufwand verringert und die Konzentration der Experten auf inhaltliche Fragen gefördert werden.

Abschließend bleibt zu bemerken, dass das Datenmodell für die Repräsentation von Data Dictionaries domänenunabhängig gestaltet ist und der *Onto-Builder* somit auch in anderen Anwendungsgebieten als dem der klinischen Studien eingesetzt werden kann.

6 Literatur

1. Cheson BD, Horning S, Coiffier B, et al: Report of an international workshop to standardize response criteria for non-hodgkin's lymphomas. J Clin Oncol 17(4):1244-1253, 1999.
2. Pschyrembel W, Dornblüth O (Hrsg.): Pschyrembel Klinisches Wörterbuch. 258. Aufl.Walter de Gruyter, Berlin, 1998.
3. Heller B, Herre H: Ontological foundations of medical information systems. Procs SoMeT 2004:3-17, 2004.
4. Rector AL: Clinical terminology: Why is it so hard? Methods Inf Med 38(4-5):239-252, 1999.
5. http://www.onto-builder.de
6. http://www.onto-med.de
7. Heller B, Herre H, Lippoldt K, Löffler M: Standardized terminology for clinical trial protocols based on ontological top-level categories. Procs CGP 2004:46-60, 2004.
8. Heller B, Herre H, Lippoldt K: The theory of top-level ontological mappings and its application to clinical trial protocols. Procs EKAW 2004:1-14, 2004.
9. http://www.sop-creator.de/
10. http://www.oncoworkstation.de/