# Automatische Erfassung verteilter RDF Beschreibungen von Klinischer Studien mit GRDDL

Dietzold S[1], Mücke R[2]
[1]*Institut für Informatik, Universität Leipzig, Deutschland*
[2]*Institut für Medizinische Informatik, Statistik und Epidemiologie, Universität Leipzig, Deutschland*
*dietzold@informatik.uni-leipzig.de, roland.muecke@imise.uni-leipzig.de*

**Einleitung und Fragestellung** Informationen zu klinischen Studien werden im World Wide Web dezentral, nicht-standardisiert, durch viele Organisationen und auf vielen verschiedenen Webseiten veröffentlicht. Studienverzeichnisse und Studienregister wollen einen zentralen Einstiegspunkt mit einer spezialisierten Suche zu klinischen Studien für Ärzte, Patienten und Wissenschaftler bieten. Die Akzeptanz einer übergeordneten Instanz für klinische Studien ist bei den durchführenden Organisationen und Personen aber nicht gegeben, da nur wenige ihre Informationshoheit abgeben und bei der Veröffentlichung von Studieninformationen abhängig sein wollen. Eine Indexierung aller vorhandenen Dokumente über einen Web-Crawler ist als Alternative zu einem zentralen Studienregister zwar möglich, eine Filterung von speziellen Informationen zu Studien und insbesondere eine semantische Analyse der Dokumente aber schwierig. Die parallele Veröffentlichung von semantisch eindeutigen Metadaten im Resource Description Framework (RDF, [1]) auf der Basis von OWL Ontologien oder RDF Schemata [2, 3] löst das Problem der semantischen Analyse der indexierten Daten und macht verteilte Informationen von klinischen Studien dadurch suchbar. Nachteil dieses Ansatzes ist aber die parallele Publikation mehrerer Dokumente mit teilweise gleichem Inhalt.
Die hier beschriebene Methode bietet die Möglichkeit, dezentrale Informationen zu Klinischen Studien zu Erfassen und daraus semantisch klare RDF Inhalte zu generieren. Dies kann durchgeführt werden, ohne dass diese Daten ein zweites Mal im RDF generiert und veröffentlicht werden müssen und ohne ein zentrales Register oder Verzeichnis mit einbeziehen zu müssen.
**Material und Methoden** Gleaning Resource Descriptions from Dialects of Languages (GRDDL, [4]) ist eine generelle Vorgehensweise mit welcher RDF Inhalte aus existierenden XML- und XHTML-Dokumenten extrahiert werden können. Die Methode nutzt das link-Element im Kopf der Dokumente für einen Verweis auf ein XSLT- Dokument [5]. Dieses Dokument beschreibt eine Transformation, welche auf bestimmten Elementen des Quell-Dokumentes aufsetzt und deren Inhalte oder Attribute in RDF umsetzt. Eine einzelne Transformation erwartet dabei so wenig wie möglich vorgegebene Struktur innerhalb des Quell-Dokumentes und arbeitet dafür z. B. mit class-Attributen. Dies ermöglichst eine einfache Integration von Transformations-Ankern in bestehende Dokumente. Der transformierende Prozess benötigt mit dieser Methode weder Informationen über zugrunde liegende Ontologien und Regeln noch ein eigenes Archiv an passenden Transformationen für verschiedene Dokumente. Er wendet die Methode an und erhält ein RDF Modell.
**Ergebnisse** Auf Basis der Website des Kompetenznetzes Maligne Lymphome (http://www.lymphom.de) haben wir eine Transformation entwickelt, welche die öffentlich zugänglichen Informationen zu Studien in das RDF transformiert. Das resultierende RDF-Modell nutzt sowohl Vokabular aus Dublin Core [6] als auch aus dem FOAF Projekt [7] und bildet allgemeine Metadaten und assoziierte PDF-Dokumente ab. Für die Durchführung des Experiments wurde ein minimaler Web-Crawler mit Hilfe des Redland RDF Frameworks [8] entwickelt, welcher als Ergebnis das extrahierte RDF-Modell liefert. In einem weiteren Schritt wurden die extrahierten Studiendaten der Suchmaschine Google Base [9] in der Kategorie „Clinical Trials“ übergeben und damit eine Suche von Klinischen Studien nach spezifischen Parametern (Phase, Condition, etc.) realisiert (siehe Abbildung).

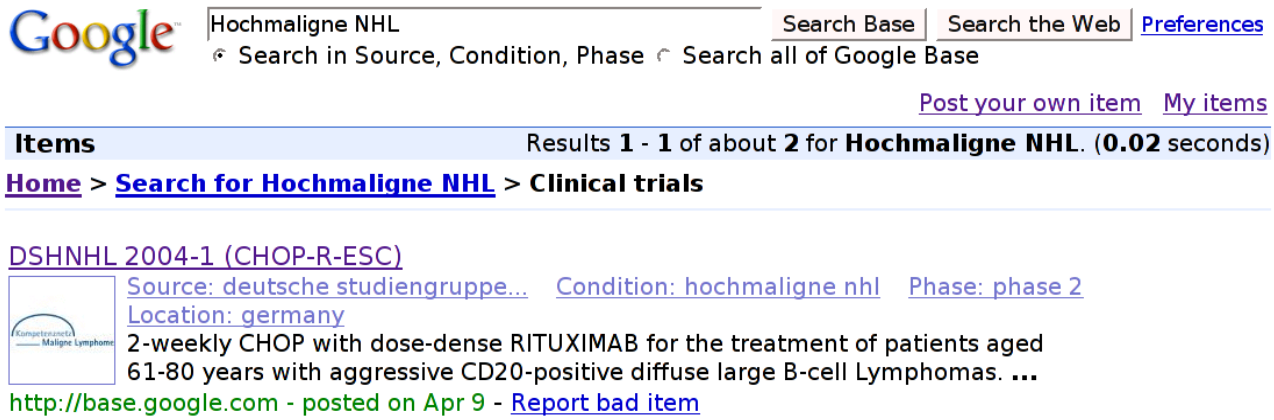


Abbildung 0: Screenshot von Google Base mit Informationen zur DSHNHL-Studie 2004-1

**Diskussion und Ausblick** Die in diesem Beitrag vorgeführte Methode extrahiert Metadaten über Klinische Studien ins RDF und ermöglicht eine Suche nach spezifischen Parameter über die Suchmaschine Google Base. Dabei werden jedoch sehr allgemeine RDF Schemata für die Beschreibung verwendet. Diese sollten in Zukunft verfeinert und erweitert werden um detailliertere Metadaten für Klinische Studien abbilden zu können (Ein- & Ausschluss-Kriterien, etc.).

## Literatur

[1] Lassila, O. & Swick, R.R. Resource Description Framework (RDF) Model and Syntax Specification: World Wide Web Consortium (W3C), 1999
[2] Schneider, P. F. P.; Hayes, P. & Horrocks, I. OWL Web Ontology Language - Semantics and Abstract Syntax: World Wide Web Consortium (W3C), 2004
[3] Brickley, D. & Guha, R. V. RDF Vocabulary Description Language 1.0 - RDF Schema: World Wide Web Consortium (W3C), 2004
[4] Hazaël-Massieux, D. & Connolly, D. Gleaning Resource Descriptions from Dialects of Languages (GRDDL): World Wide Web Consortium (W3C), 2005
[5] Clark, J. XSL Transformations (XSLT): World Wide Web Consortium (W3C), 1999
[6] Sugimoto, S.; Baker, T. & Weibel, Dublin Core: Process and Principles. In: Lim, E.; Foo, S.; Khoo, C.S.G.; Chen, H.; Fox, E.A.; Urs, S.R. & Thanos, C. (Hrsg.) Digital Libraries: People, Knowledge, and Technology, 5th International Conference on Asian Digital Libraries, ICADL 2002 Singapore, December 11-14, 2002, Proceedings, Springer, 2002, 2555, 25-35
[7] Brickley, D. & Miller, L. FOAF Vocabulary Specification - http://xmlns.com/foaf/0.1/: FOAF Project, 2004
[8] Beckett, D. The design and implementation of the Redland RDF Application Framework: Computer Networks, 2002, 39, 577-588
[9] Google Base Suche nach “Clinical Trials”: http://base.google.com/base/search?a_n0=clinical+trials&scoring=ld&a_y0=9