André Langer

SemProj - Ein Semantic Web-basiertes System zur Unterstützung von Workflow- und Projektmanagement

Diplomarbeit

Diplom.de

Bibliografische Information der Deutschen Nationalbibliothek:

Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de/ abrufbar.

ISBN: 9783836611350

André Langer

SemProj - Ein Semantic Web-basiertes System zur Unterstützung von Workflow- und Projektmanagement

Diplom.de

André Langer

SemProj –
Ein Semantic Web-basiertes System zur Unterstützung von Workflow- und Projektmanagement

Diplomarbeit
Technische Universität Chemnitz
Fakultät für Informatik
Studiengang Angewandte Informatik
November 2007

Diplomica GmbH
Hermannstal 119k
22119 Hamburg

Fon: 040 / 655 99 20
Fax: 040 / 655 99 222

agentur@diplom.de
www.diplom.de

André Langer
SemProj - Ein Semantic Web-basiertes System zur Unterstützung von Workflow- und Projektmanagement

ISBN: 978-3-8366-1135-0
Druck Diplomica® Verlag GmbH, Hamburg, 2008
Zugl. Technische Universität Chemnitz, Chemnitz, Deutschland, Diplomarbeit, 2007

Zusammenfassung

Mit mehr als 120 Millionen registrierten Internetadressen (Stand: März 2007) symbolisiert das Internet heutzutage das größte Informationsmedium unserer Zeit. Täglich wächst das Internet um eine unüberschaubare Menge an Informationen. Diese Informationen sind häufig in Dokumenten hinterlegt, welche zur Auszeichnung die Hypertext Markup Language verwenden. Seit Beginn der Neunziger Jahre hat sich dieses System bewährt, da dadurch der einzelne Nutzer in die Lage versetzt wird, auf einfache und effiziente Weise Dokumentinhalte mit Darstellungsanweisungen zu versehen und diese eigenständig im Internet zu veröffentlichen. Diese Layoutinformationen können bei Abruf der entsprechenden Ressource durch ein Computerprogramm leicht ausgewertet und zur Darstellung der Inhalte genutzt werden. Obwohl sowohl die Layoutinformationen als auch die eigentlichen Dokumentinhalte in einem textuellen Format vorliegen, konnten die Nutzertextinhalte durch eine Maschine bisher nur sehr eingeschränkt verarbeitet werden. Während es menschlichen Nutzern keinerlei Probleme bereitet, die Bedeutung einzelner Texte auf einer Webseite zu identifizieren, stellen diese für einen Rechner prinzipiell nur eine Aneinanderreihung von ASCII-Zeichen dar.

Sobald es möglich werden würde, die Bedeutung von Informationen durch ein Computerprogramm effizient zu erfassen und weiterzuverarbeiten, wären völlig neue Anwendungen mit qualitativ hochwertigeren Ergebnissen im weltweiten Datennetz möglich. Nutzer könnten Anfragen an spezielle Agenten stellen, welche sich selbstständig auf die Suche nach passenden Resultaten begeben; Informationen verschiedener Informationsquellen könnten nicht nur auf semantischer Ebene verknüpft, sondern daraus sogar neue, nicht explizit enthaltene Informationen abgeleitet werden. Ansätze dazu, wie Dokumente mit semantischen Metadaten versehen werden können, gibt es bereits seit einiger Zeit. Lange umfasste dies jedoch die redundante Bereitstellung der Informationen in einem eigenen Dokumentenformat, weswegen sich keines der Konzepte bis in den Privatbereich durchsetzen konnte und als Endkonsequenz in den vergangenen Monaten besonderes Forschungsinteresse darin aufkam, Möglichkeiten zu finden, wie semantische Informationen ohne großen Zusatzaufwand direkt in bestehende HTML-Dokumente eingebettet werden können.

Im Rahmen der vorliegenden Publikation werden diese neuen Möglichkeiten im Bereich des kollaborativen Arbeitens näher untersucht. Ziel ist es dazu, eine Webapplikation zur Abwicklung typischer Projektmanagement-Aufgaben zu entwickeln, welche jegliche Informationen unter einem semantischen Gesichtspunkt analysieren, aufbereiten und weiterverarbeiten kann und unabhängig von der konkreten Anwendungsdomain und Plattform systemübergreifend eingesetzt werden kann. Die Konzepte Microformats und RDFa werden dabei besonders herausgestellt und nach Schwächen und zukünftigen Potentialen hin analysiert.

Abstract

The World Wide Web supposably symbolizes with currently more than 120 million registered internet domains (March 2007) the most comprehensive information reference of all times. The amount of information available increases by a storming bulk of data ever day. Those information is often embedded in documents which utilize the Hypertext Markup Language. This enables the user to mark out certain layout properties of a text in an easy and efficient fashion and to publish the final document containing both layout and data information. A computer application is then able to extract style information from the document resource and to use it in order to render the resulting website. Although layout information and data are both equally represented in a textual manner, a machine was hardly capable of processing user content so far. Whereas human consumers have no problem to identify and understand the sense of several paragraphs on a website, they basically represent only a concatenation of ASCII characters for a machine.

If it were possible to efficiently disclose the sense of a word or phrase to a computer program in order to process it, new astounding applications with output results of high quality would be possible. Users could create queries for specialized agents which autonomously start to search the web for adequate result matches. Moreover, the data of multiple information sources could be linked and processed together on a semantic level so that above all new, not explicitly stated information could be inferred. Approaches already exist, how documents could be enhanced with semantic meta-data, however, many of these involve the redundant provision of those information in a specialized document format. As a consequence none of these concepts succeeded in becoming a widely used method and research started again to find possibilities how to embed semantic annotations without huge additional efforts in an ordinary HTML document.

The present publication focuses on an analysis of these new concepts and possibilities in the area of collaborative work. The objective is to develop the prototype of a web application with which it is possible to manage typical challenges in the realm of project and workflow management. Any information available should be processable under a semantic viewpoint which includes analysis, conditioning and reuse independently from a specific application domain and a certain system platform. Microformats and RDFa are two of those relatively new concepts which enable an application to extract semantic information from a document resource and are therefore particularly exposed and compared with respect to advantages and disadvantages in the context of a “Semantic Web”.

Danksagung

"Whatever you do will be unimportant, but it is very important that you do it." ~ Mahatma Gandhi

An dem vorliegenden Buch haben viele Personen einen direkten oder indirekten Einfluss gehabt, denen ich diesen kurzen Abschnitt widmen möchte. Die Aufgabe, eine erste eigene Publikation herauszugeben, scheint zu Beginn lang und das gewählte Thema scheint das einzig Wichtige zu sein, was den Inhalt der Arbeit bestimmt. Ich persönlich habe während der Anfertigung des Buches Erfahrungen gemacht, die über diesen inhaltlichen Fokus weit hinausgehen. Sicherlich stand zu Beginn die Frage nach einiger geeigneten Thematik im Raum; und ich bin froh dieses sehr zukunftsträchtige Thema gewählt zu haben, da ich mich sonst sicherlich nicht derart intensiv damit auseinander gesetzt hätte. Viel interessanter war für mich jedoch zu sehen, wie ich mich persönlich weiterentwickelt habe. Die Anfertigung war dabei ein dynamischer Prozess, der durch Erfolge aber auch durch kleine und große Probleme geprägt war, die zu Beginn nicht immer abzusehen, aber letztendlich alle überwunden werden konnten. Ich habe im Zuge dessen neue, interessante Methoden, Ansätze und Kontakte kennen gelernt, die im späteren Berufsleben von großem Nutzen sein können.

An erster Stelle möchte ich mich dazu bei Prof. Dr. Martin Gaedke von der Technischen Universität Chemnitz bedanken, der wesentlichen Anteil an der gemeinsamen Entwicklung des Themas hatte und mehrfach seinen Feierabend entfallen ließ, um mit mir stundenlang aktuelle Ergebnisse bis in die Nacht zu diskutieren. Außerdem danke ich dem gesamten Forschungsteam der Professur Verteilte und Selbstorganisierende Rechnersysteme, welche sich in der zurückliegenden Zeit bemerkenswert einsetzten, um mit neuen Lehrmethoden alle Studierenden dazu zu ermutigen, über den Tellerrand hinauszuschauen und die zugrunde liegenden Zusammenhänge zu begreifen, um diese in der Praxis später eigenständig anwenden zu können.

Mein weiterer Dank gilt meinen Großeltern, welche mich in der Anfertigungzeit mit allen nur erdenklichen Hilfen unterstützten. Ebenso möchte ich mich bei meinen Freunden bedanken, die mich immer wieder ermutigt und motiviert haben, mein Ziel zu verfolgen. Insbesondere seien hier Maja Heidrich, bei der ich mich für das Korrekturlesen der Studie bedanken möchte, sowie Thomas Fichtner genannt, der mir immer wieder Feedback bei der Implementierung des Prototypen gab.

Mein abschließender Dank gilt dem Unternehmen der 09111 Studio Chemnitz GmbH&Co. KG, welche auch in den letzten zwei Monaten eine sehr flexible Arbeitszeitgestaltung ermöglichten sowie der Stiftung der Deutschen Wirtschaft, bei der ich mich für die Förderung im Rahmen des Stipendiatenprogramms in den zurückliegenden vier Jahren und das entgegengebrachte Vertrauen bedanken möchte.

Die Thematik in dem vorliegenden Dokument habe ich versucht, von verschiedenen Blickrichtungen so zu bearbeiten, sodass sich ein geschlossenes Gesamtbild ergibt, welches auch für Unbeteiligte interessant und gut verständlich ist. Wenn das vorliegende Buch in Zukunft auch für andere Personen und Institutionen von Nutzen ist und zu neuen Erkenntnissen beiträgt, würde ich mich freuen.

André Langer im November 2007

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AJAX	Asynchronous Javascript and XML
API	Application Programming Interface
BPMN	Business Process Modeling Notation
BSD	Berkeley Software Distribution
DAML	DARPA Agent Markup Language
DC	Dublin Core
DTD	Document Type Definition
EPK	Ereignisgesteuerte Prozesskette
FOAF	Friend of a Friend
GPL	General Public License
GRDDL	Gleaning Resource Descriptions from Dialects of Languages
HTML	Hypertext Markup Language
NS	Namespace
OIL	Ontology Inference Layer
OWL	Web Ontology Language
RDF	Ressource Description Framework
RDFS	RDF Schema
SGML	American Standard Code for Information Interchange
SHOE	Simple HTML Ontology Extensions
UML	Unified Modeling Language
URI	Uniform Ressource Identifier
URL	Uniform Ressource Locator
W3C	World Wide Web Consortium
WfMC	Workflow Management Coalition
WFMS	Workflow-Management-System
XHTML	eXtensible Hypertext Markup Language
XML	eXtensible Markup Language
XPDL	XML Process Definition Language
XSD	XML Schema Definition
XSL	eXtensible Stylesheet Language
XSLT	eXtensible Stylesheet Language Transformations

1. Die Vision des Semantic Web

"Things are only impossible until they're not." ~ Star Trek: The Next Generation - When The Bough Breaks, *Jean-Luc Picard, Season 1, 1988*

1.1. Motivation

Ein junger Unternehmensberater sitze an einem Dienstagmorgen in der Cafeteria eines namhaften Hotels irgendwo im Herzen von Deutschland. Während er mit einem Löffel in der linken Hand die Tasse Kaffee vor sich auf dem Tisch umrührt, hält er in der rechten Hand bereits seinen PDA, um sich einen Überblick über die Termine des kommenden Tages zu verschaffen. Nach einem Klick auf „*heutige Termine anzeigen*" liefert das System die Beschreibung: „*Sie haben heute um 10.00 Uhr ein Treffen mit Dr. Müller in der Ulmenstraße 5. Um 15.30 Uhr ist eine Telefonkonferenz mit der Arbeitsgruppe Asterisk angesetzt. 17.00 Uhr haben Sie einen Termin mit Herrn Holger Schmidt vorgemerkt. Um 18.20 Uhr schließlich startet ihr Flug nach Grenoble.*" Der Berater überlegt, dass es besser wäre, den Termin mit *Herrn Schmidt* zu verlegen, um rechtzeitig auf dem Flughafen zu sein. Er verschiebt den Termin um zwei Tage, was durch das System zunächst abgelehnt wird mit der Begründung, dass sich *Herr Schmidt* zu dieser Zeit auf einer Dienstreise befindet. Nach einer Bestätigung legt das System das Treffen auf den nächstmöglichen Termin und versendet an *Herrn Schmidts* Emailadresse eine entsprechende Nachricht. Die nächste Anfrage, worum es bei der Telefonkonferenz geht, beantwortet das System mit einer Meldung „*Die Firma Meier möchte diskutieren, was ein Umstieg auf SCCP für Vorteile bringt.*" *SCCP* ist dem Berater unbekannt, worauf ihm das System den Begriff definiert[1] und weitere Informationen liefert. Nach dem Frühstück macht sich der Berater auf zu seinem ersten Termin, lässt von dem System automatisch ein *Taxi* rufen und fragt auf dem Weg nach draußen noch schnell ab, wann und wo er *Dr. Müller* letztmalig getroffen und wo sein Geschäftspartner den letzten Urlaub verbracht hat.

Das skizzierte Szenario scheint auf dem ersten Blick in ferner Zukunft zu liegen. Derartige Interaktionen zwischen Mensch und Maschine werden häufig im Bereich der Science-Fiction angesiedelt. Zu allgegenwärtig scheinen eigene Erfahrungen, wie aufwändig es sein kann, eine Suche nach spezifischen Daten im Internet mit Computerunterstützung durchzuführen. Jegliche Informationen, welche nicht explizit in einer Organizer-Software gespeichert sind und entsprechend angezeigt werden können, scheinen für einen Rechner nicht auswertbar. Zu groß wäre der Suchaufwand in der Informationsfülle des heutigen Internets – ganz abgesehen von der Fragestellung, wie ein Rechner Beziehungen zwischen Informationen herstellen solle, Mehrdeutigkeiten auflösen und eine auf die Fragestellung zugeschnittene, für den Menschen verständliche und vereinfachte Ausgabe zurückliefern kann.

[1] Skinny Client Control Protocol, ein von Cisco Systems Inc. entwickeltes, proprietäres Protokoll zur Abhaltung von Telefonkonferenzen über VoIP in Echtzeit

Dennoch ist eine derartige Anwendung heutzutage vorstellbar und nicht länger Fiktion. Der Traum von intelligenten Maschinen besteht schon seit langer Zeit. Beschränkt man sich nur auf Entwicklungen mit Bezug zur modernen Rechentechnik, so prägte Alan Turing 1950 erstmals die Vorstellung von intelligenten Maschinen [Tur50]. Der sich daran anschließende Optimismus unter KI-Forschung bis Ende der 60er Jahre wurde durch eine Phase der Ernüchterung beendet, als man zunehmend die Beschränkungen grundlegender Konzepte und der darauf basierenden Algorithmen erkannte. Während in den Siebziger und Achtziger Jahren neue Erfolge bei der Entwicklung kommerzieller Expertensysteme erzielt worden, lag der Fokus bei der Verbreitung des World Wide Webs nach Einführung des Hypertext-Konzepts durch Sir Tim Berners-Lee im Jahr 1989 zunächst auf der öffentlichen Bereitstellung von Informationen in Dokumenten, welche auch für Endanwender einfach zu realisieren sein sollte. Die Hypertext Markup Language (HTML) als Anwendung der Standard Generalized Markup Language (SGML), mit der Dokumentstrukturen syntaktisch beschrieben werden konnten, schuf dabei die Grundlage für einen Dokumenttyp, über den mithilfe einiger zusätzlicher Auszeichnungen zur Angabe von Textformatierungen alle Dokumentinformationen in einem unitären Dokumentformat in einer einheitlichen Struktur gespeichert, verbreitet und zur Darstellung bereitgestellt werden konnten. Nutzer wurden dadurch in die Lage versetzt, Textinhalte einfach formatieren und mit anderen Dokumenten und Multimediadateien verknüpfen zu können. Verbunden mit der einfachen Erlernbarkeit und Einsetzbarkeit von HTML durch Privatanwender war jedoch die Einschränkung, einen festen Satz an vordefinierten Auszeichnungselementen („Tags“) vorzugeben, welchen eine explizite Bedeutung zugeordnet wurde. Im Bezug auf die Rücktransformation von Dokumentinhalten in Informationsstrukturen stellte dies ein Problem dar, da Informationen über die Natur der dargestellten Inhalte nicht gespeichert wurden (es konnten bzw. sollten keine eigenen Auszeichnungselemente hinzugefügt werden), sondern HTML in erster Linie die Dokumentdarstellung unterstrich. Durch den einfachen Syntax und die schnelle Bereitstellung von Entwicklungswerkzeugen (WYSIWYG-Editoren) verbreitete sich das Hypertext-Konzept binnen kurzer Zeit und fand breite Akzeptanz[2], dennoch wurden besonders im Bereich der rechnergestützten Dokumentverarbeitung Anforderungen immer wichtiger, nicht nur Darstellungsinformationen sondern auch syntaktische Dokumentstrukturinformationen in einem universellen, einheitlichen Format speichern und übertragen zu können. Als Folge davon wurde die eXtensible Markup Language (XML) entwickelt, welche sich bis heute als Austauschformat in unterschiedlichsten Anwendungsbereichen etabliert hat.

[2] An manchen Stellen wird in HTML dadurch sogar der wohl bedeutendste Schritt gesehen, der erst das Internet in der heutigen Realisation ermöglichte, da die Auszeichnungssprache weltweit akzeptiert wurde [Lac05, p. 1]. Vielfach wurde versucht, weitere Standardsprachen („lingua franca des Internets“) für andere Anwendungsbereiche einzuführen, die jedoch alle mehr oder weniger an unterschiedlichen Auffassungen verschiedener Organisationen und Länder gescheitert sind oder modifiziert werden mussten.

HTML und XML (und weitere Konzepte wie CSS) widersprechen dabei nicht einander, sondern können sich gegenseitig ergänzen und ineinander überführt werden (XSLT) , da beide der gleichen Sprachfamilie entstammen und mit der Spezifikation von XHTML1.0 dies sogar eine Untermenge von XML mit einem dedizierten Anwendungsfeld darstellt. Je nach Anwendungsfall mit Schwerpunkt auf Layout, Struktur oder Inhalt wird man sich für eine konkrete Umsetzungsmöglichkeit entscheiden, wodurch eine gewisse Trennung in einen layoutorientierten oder strukturorientierten Ansatz vorhanden ist. Trotz dass XML im letzteren Fall schon den Anschein erweckt, Informationen über Dokumentinhalte perfekt abbilden zu können, reichte aber auch dies nicht aus, um Inhalte für einen Computer verständlich werden zu lassen, da damit zwar eine syntaktische Korrektheit sichergestellt werden konnte, nicht jedoch eine semantische Korrektheit.[3]

1.2. Beschränkungen des heutigen Internets

Das World Wide Web ist im Wesentlichen ein Medium zur Veröffentlichung und Distribution von Informationen. Ende Dezember 2006 waren weltweit 120 Millionen Domains registriert. Pro Monat kommen statistisch 3,2 Millionen neue Domains hinzu. [Ver07]. Die Anzahl der tatsächlich im Internet vorhandenen Dokumente kann nur geschätzt werden, wobei Millionen privater Homepages mit wenigen Unterseiten und Webseiten großer Unternehmen mit womöglich mehr als 100.000 Dokumenten gleichermaßen ins Gewicht fallen. Eine Hochrechnung basierend auf Daten des Archivierungsdienstes www.archive.org im Frühjahr 2006 geht von einer kumulativen Summe von insgesamt 55 Milliarden (1 Petabyte) Dokumenten seit 1996 aus mit einem Zuwachs von 20 Terabyte an Daten pro Monat. Andere Statistiken sehen diese Schätzungen als zu pessimistisch an und sprechen von einer zwei Drittel höheren Dokumentenanzahl. [Wen06] Diese Zahl von über 160 Milliarden Dokumenten im World Wide Web basiert auf der Annahme, dass aktuelle Suchmaschinen real nur ca. 35 % aller tatsächlich vorhandenen Dokumente indiziert haben (Abbildung 1). Im Umkehrschluss bedeutet dies, dass ein Großteil potentiell relevanter Informationen entweder gar nicht oder nur indirekt gefunden werden kann. Selbst die Informationen, auf die durch Suchmaschinen öffentlich zugegriffen werden kann, existieren unabhängig voneinander. Abgesehen von Techniken wie Web Scraping oder explizit angebotenen Web Services ist der Datenbestand einer Webapplikation auf die eigene Anwendung bzw. Domain begrenzt, über die der zuständige Entwickler zusätzliches Hintergrundwissen besitzt. Könnte man den Inhalt der Milliarden Dokumente miteinander verknüpfen, daraus Informationen abrufen, den Inhalt und Wahrheitswert überprüfen und mitunter sogar neue Informationen ableiten, wären ein Informationsgewinn und Anwendungen denkbar, die derzeit noch nicht vorstellbar sind, da das spezifische Suchen, Vergleichen und Ableiten von Informationen bisher menschliche Interaktion erfordert.

[3] Breitman, Casanova und Truszkowski bezeichnen in diesem Zusammenhang das heutige Internet als „Syntactic Web“ im Gegensatz zum angestrebten „Semantic Web“ [Bre07, p.5]

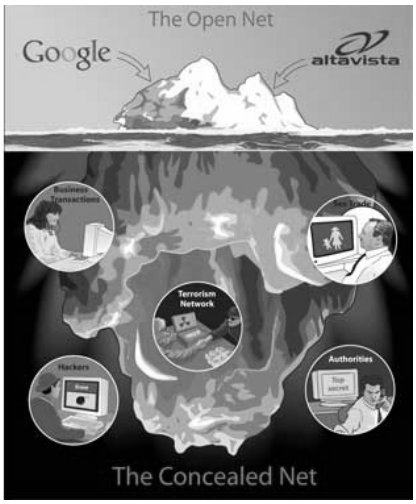

Abbildung 1: Die Sicht einer heutigen Suchmaschine auf das WWW, Quelle: http://www.foi.se/FOI/templates/Page____4070.aspx

Ein weiteres Problem stellt darauf aufbauend die Art und Weise der Suche nach Informationen dar. Moderne Algorithmen, die in heutigen Suchmaschinen zur Anwendung kommen, sind zwar hochkomplex und haben sich in den vergangenen zwei Jahrzehnten qualitativ kontinuierlich weiterentwickelt, doch ist das Grundkonzept dahingehend gleich geblieben, dass alle Suchanfragen in der Regel auf einzelnen Schlüsselwörtern basieren und deren Vorkommen im (Volltext-)Index des Suchmaschinendatenbestandes überprüft wird. Die Konsequenz daraus ist, dass zwar syntaktische Analysen durchgeführt können und nach ähnlichen Wörtern gesucht werden kann, in aller Regel auch die Anzahl des gemeinsamen Vorkommens der Schlüsselwörter im Zieldokument sowie die Popularität der Seite eine Rolle spielt, doch bleibt die Bedeutung der Anfrage in aller Regel verborgen. So ist aus einem einzelnen Schlüsselwort wie etwa *Flügel* nicht ableitbar, ob der Nutzer *eine Webseite sucht, die Musikinstrumente anbietet, eine Anleitung für eine mechanische Konstruktion sucht,* oder *in ein zoologisches Lexikon schauen* möchte. Dementsprechend bieten heutige Suchmaschinen in einer Übersicht häufig die „geeignetsten" Treffer an, worunter derartige Begriffsdeutungen zumeist miteinander vermischt sind. In manchen Situationen hilft es, mehrere Schlüsselwörter in die Suchanfrage aufzunehmen (*+wie +spielt +man +auf +einem +Flügel*), wodurch Dokumente gefunden werden könnten, die die Begriffe *spielen* und *Flügel* enthalten, doch wäre es intuitiver, die Frage in dieser Form direkt an einen Suchdienst stellen zu können, der diese Frage als Ganzes versteht und nur passende Resultate zurückliefert (auch solche, in denen das Wort *Flügel* nie erwähnt wird, dafür aber vielleicht von einem *Klavier* gesprochen wird).

Dass dies technisch nicht trivial umzusetzen ist, liegt wie in Abschnitt 1.1. bereits kurz beschrieben darin begründet, dass Dokumentinhalte im World Wide Web größtenteils im (X)HTML – Format vorliegen. Der Fokus auf einer einfachen Layoutbeschreibung brachte jedoch den Nachteil mit sich, dass die Beschreibung des Inhalts von HTML-Dokumenten nur zweitrangig betont wurde (Meta-Tagging). Die Folge davon ist, dass die entstandenen HTML-Dokumente zwar in ihrer Struktur durch einen Computer gut verarbeitbar sind, der Inhalt computergeschützt aber nicht trivial Konzepten der realen Welt zuzuordnen ist (die Tags zur Auszeichnung können entfernt werden, der resultierende Text an sich ist strukturlos und kann nur mithilfe von String-Operationen verarbeitet werden), was wiederum zu dem angesprochenen Problem führt, bestimmte Sachverhalte bei einem Indexierungsvorgang semantisch voneinander unterscheiden zu können.