Bachelorarbeit

Konzeption und prototypische Implementierung eines generischen Klassifizierungswerkzeuges für ein Dokumentenmanagement-System auf Basis von IBM Lotus Notes

vorgelegt bei

Prof. Dr. Ludwig Nastansky

betreut durch

Dipl.-Wirt.-Inf. Bernd Hesse

Sommersemester 2007

vorgelegt von

Michael Suren

Student der Wirtschaftsinformatik

Matrikelnummer 6219620

Kirchweg 23, 33181 Bad Wünnenberg - Haaren

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

DTD	Document Type Definition
GCC	Groupware Competence Center
IBM	International Business Machines Corporation
K-Pool	Knowledge Pool
RDF	Ressource Description Framework
URI	Uniform Ressource Identifier
XML	Extensible Markup Language

Glossar

agglomerativ	Schrittweise Zusammenfassung von Objekten zu Gruppen
divisiv	Schrittweise Aufteilung einer Gesamtheit in Gruppen
Dokument	Gegenstände, die physisch einschließlich digital vorhanden sind und Sinn und Bedeutung haben. Außer Texten gehören hierzu auch Bild- und Tondokumente.
K-Pool	Der K-Pool ist eine Wissens- und Dokumentenmanagement-umgebung auf Basis von IBM Lotus Notes Domino
Listenfelder	Felder in Lotus Notes Domino, die mehrere Werte aufnehmen können.
Semantik	Bedeutung eines Wortes / Ausdruckes
Tag	Ein Tag ist ein Schlagwort
Term	Ein Begriff.
URI	Uniform Ressource Identifier ist eine Zeichenkette, die zur Identifizierung einer Ressource dient.

1 Einleitung

1.1 Thematische Einführung

„Eine der wichtigsten Fähigkeiten, die zu lernen ist in einer Welt

mit Informationsüberflutung, ist Filtern und Suchen.“

-Herwart Holland-Moritz (1951 – 2001),

Gründungsmitglied Chaos Computer Club

Im Zeitalter moderner Kommunikations- und Datenbanksysteme sowie des World Wide Web stehen dem einzelnen Menschen unzählig viele Informationsquellen zur Verfügung. Diese Quellen bieten eine Fülle an Daten und Informationen, doch das Auffinden von Wissen wird in diesen Datenmengen fast zu einem unmöglichen Unterfangen.

In vielen Unternehmen verhält es sich ähnlich. Im Laufe der Zeit sammeln sich Unmengen an Informationen an, auf deren Basis die unternehmensinternen Prozesse ablaufen und die dem Unternehmen helfen können, sich im Wettbewerb durchzusetzen. Diese Informationen, die zum Unternehmenserfolg beitragen und eine Art Wissensbasis bilden, gehen allerdings oftmals in den Systemen des Unternehmens unter. In vielen Fällen weiß das Unternehmen gar nicht, über welches Wissen es eigentlich verfügt. Es existiert keine gemeinsame Informations- und somit auch keine Wissensbasis.

Die Schwierigkeit, eine Wissensbasis im Unternehmen aufzubauen, wird in verschiedenen Bereichen deutlich. Zum einen gibt es das Fachwissen, welches in den einzelnen Köpfen der Mitarbeiter existiert und mit deren Ausscheiden aus dem Unternehmen verloren geht. Oftmals ist die Personalabteilung des Unternehmens die einzige Stelle, die annähernd Kenntnis über die Fähigkeiten des jeweiligen Mitarbeiters hat. Auf der anderen Seite gibt es eine enorme Informationsfülle, die sich z.B. in den Maildatenbanken eines jeden Mitarbeiters anhäuft. Da der einzelne Mitarbeiter Informationen in seiner persönlichen Maildatenbank ansammelt und aufbewahrt, wird diese somit häufig zu einer Informationsdatenbank umfunktioniert. In der Maildatenbank kann aber nur sehr grob z.B. nach Name, Datum oder Dateigröße sortiert werden. An dieser Stelle gezielt Informationen wiederzufinden stellt sich als sehr schwierig heraus.

Dem Menschen helfen bei einer solchen Gewinnung von Informationen Kategorisierungs-Systeme, da der Mensch von Natur aus in „Schubladen“ denkt. Diese Kategorien

helfen beim Navigieren in vorhandenen Informationen, da sie die Möglichkeit eröffnen, sich einer gesuchten Information schrittweise anzunähern. Ein Beispiel für eine solche schrittweise Annäherung ist die Beschreibung der Lage des Wohnortes. Menschen leben in Häusern, Straßen, Dörfern, Städten, Kreisen, Bundesländern, Ländern, Kontinenten. Soll jemandem nun der Ort beschrieben werden, an dem man lebt, so kann man sich anhand dieser hierarchischen Kategorien Stufe für Stufe dem Ziel nähern. Auf die gleiche Art und Weise kann Wissen kategorisiert werden.

Diese Kategorisierung ist der erste Schritt, die Informationsflut der heutigen Arbeitswelt, die täglich auf eine Person einwirkt, bewältigen zu können. Zudem bieten Kategorisierungen eine Übersicht über vorhandene Wissensbestände an. Ein bekanntes Beispiel sind Webkataloge im World Wide Web wie Yahoo![1].

1.2 Zielsetzung

Ziel dieser Arbeit ist die Konzeption und Implementierung eines Kategorisierungs-Systems, das dem Anwender eine Hilfestellung gibt, Dokumente zu kategorisieren und vorhandene Kategorien von Dokumenten zu ändern. Dieses System soll zudem in Datenbanken auf Basis von Lotus Notes Domino eingesetzt werden können.

1.3 Aufbau der Arbeit

Auf Basis der oben beschriebenen Problemstellung werden zu Beginn dieser Arbeit in Kapitel 2 die unterschiedlichen Kategorisierungsmethoden vorgestellt, um die Möglichkeiten sowie Arbeitsweisen der jeweiligen Methoden darzustellen. Im Anschluss werden in Kapitel 3 die Kategorisierungsverfahren analysiert, die bereits im Groupware und Knowledge Management System des GCC eingesetzt werden, um Informationen gezielt einzuordnen und wiederzufinden. Das Ergebnis dieser Analyse bildet die Grundlage für das Konzept der generischen Kategorisierung, welches in Kapitel 4 beschrieben wird. Aufbauend auf dem Konzept, das in Kapitel 4 erstellt wurde, wird in Kapitel 5 das Ergebnis der prototypischen Implementierung beschrieben. In Kapitel 6 wird ein Ausblick gegeben, der weitere Ansätze für mögliche zukünftige Verbesserungen liefert. Zum Ende der Arbeit werden in Kapitel 7 die Ergebnisse in einem Fazit noch einmal kompakt zusammengefasst.

[1] www.yahoo.de

2 Thematische Grundlagen

In diesem Kapitel werden die thematischen Grundlagen erläutert, auf denen die weiteren Ausführungen dieser Arbeit beruhen.

Es werden unterschiedliche Kategorisierungsverfahren und deren mögliche Einsatzgebiete vorgestellt.

2.1 Kategorisierungsverfahren

2.1.1 Data Mining

Nach *Shmueli et al.* ist Data Mining eine noch sehr neue Methode der Informationsgewinnung**,** bei der sich Klassifikationen mit Hilfe von Neuronalen Netzen, Entscheidungsbäumen und Logarithmischer Regression erstellen lassen.

Eine allgemein und konsistente Definition von Data Mining ist: „‘Extracting usefull information from large datasets. (Hand et al. 2001)‘“[2].

Die Einsatzmöglichkeiten von Data Mining sind neben der Erstellung von Klassifikationen die Erstellung von Prognosen und die Durchführung von Assoziationsanalysen.

Klassifikation: Ein Beispiel hierfür ist ein Ring, der bei einem Juwelier mit einer Kreditkarte bezahlt wird. Die Fragestellung, die sich bei der Erstellung einer Klassifikation der Daten in diesem Fall ergibt, lautet: „Handelt es sich um eine glaubhafte Transaktion oder handelt es sich hierbei um einen Betrugsversuch?“ Die Transaktion wird aufbauend auf einer Analyse in die Kategorie: Glaubhafte Transaktion oder Betrug eingestuft. [3]

Prognosen: Prognosen helfen, mögliche zukünftige Ereignisse vorherzusagen. Eine exemplarische Fragestellung für die Erstellung von Prognosen lautet: „Welche 10 Kunden bieten uns das größte Deckungsbeitragspotential?“

Assoziationsanalyse: Die Assoziationsanalyse wird eingesetzt, um Zusammenhänge zwischen Daten zu erkennen[4]. Diese Zusammenhänge werden durch „Wenn – dann“

[2] Siehe hierzu Shmueli (2007, S.1)

[3] Vgl. Shmueli (2007, S.91f)

[4] Vgl. Shmueli (2007, S.203)

Aussagen abgebildet. Auf diese Weise können Zusammenhänge zwischen Datensätzen in Datenbanken erkannt werden. Zum Beispiel werden Diagnosen von Krankheiten aufgrund von Zusammenhängen zwischen Symptomen und einer möglichen Krankheit erstellt. Auch der tägliche Einkauf kann dieser Analyse unterzogen werden, um festzustellen, welche Produkte zusammen gekauft werden[5]. Anhand dieser Daten sind Manager von Märkten in der Lage, Produkte geschickt im Markt zu positionieren.[6]

Das Verfahren des Data Mining umfasst folgende Schritte**,** die sequentiell durchlaufen werden:

1. Einsatzmöglichkeiten: Zuerst muss das Einsatzszenario festgelegt werden, da Data Mining in vielen Bereichen eingesetzt werden kann. Für diese Einsatzmöglichkeiten muss ein Verständnis entwickelt werden.
2. Datensatz: Es muss ein Datensatz erstellt werden, der in der Analyse verwendet werden soll. Die Daten können aus mehreren Datensätzen zusammengesetzt sein, oder sie werden per Zufall aus einer großen Datenmenge ausgewählt, sodass der Datensatz im Anschluss viele Daten (1.000 bis 10.000 Stück) umfasst.
3. Datenbereinigung: Die Datenbereinigung kontrolliert vorhandene Daten auf Vollständigkeit. Sollten Werte in den Daten fehlen, so muss entschieden werden, ob diese Daten durch einen Durchschnittswert ersetzt werden sollen. Um einen konsistenten Datenbestand zu erhalten, sind leere Daten durch Mittelwerte zu ersetzen. Eine Datenreduktion wird –falls nötig– durchgeführt, um unwichtige Variablen zu entfernen. Nach der Datenreduktion erfolgt eine Aufteilung in Trainingsdaten, Überprüfungsdaten und Testdaten.
4. Selektion Aufgabe: Es muss entschieden werden, welche Aufgabe das Data Mining erfüllen soll. Zur Auswahl stehen: Regression, Klassifikation, Prognose, Clusteranalyse (siehe 2.1.2) und Abweichungserkennung.[7]

[5] Vgl. Shmueli (2007, S.203)

[6] Vgl. Shmueli (2007, S.203)

[7] Für die einzelnen Verfahren, sowie das Data Mining lege ich dem geneigten Leser das Buch „Data Mining for Business Intelligence“ nahe.

5. Auswahl des Algorithmus: Je nach Wahl der Aufgabe aus Schritt 4 wird der entsprechende Algorithmus angewendet, wobei Variationen des Algorithmus zu einer Verbesserung der Daten beitragen.
6. Training: Die letzte Einstellung der Variationen des Algorithmus wird auf den Trainingsdatensatz angewendet, um die Werte nochmals zu überprüfen.
7. Aufstellen des Modells: Aus diesen Daten wird ein Modell aufgestellt, welches nun die Daten zu verarbeiten hat und dann ein Ergebnis liefert [8]

Data Mining bietet den Vorteil, dass es automatisch durchgeführt werden kann und dabei eine sehr große Datenmenge kategorisiert. Nachteilig an dem Verfahren ist die hohe Anzahl von Datensätzen, die zur Verfügung stehen müssen, damit aussagekräftige Daten entstehen.

Vom Kategorisierungsverfahren Data Mining ausgehend wird im Anschluss die Clusteranalyse vorgestellt.

2.1.2 Clusteranalyse

Die Clusteranalyse ist ein Kategorisierungsverfahren, bei dem Gruppen (Cluster) aus Elementen, die über ähnliche Ausprägungen in den Werten verfügen, gebildet werden. Sie dient außerdem dazu, Muster zu erkennen. Dieses Verfahren wurde schon erfolgreich in der Astronomie, Medizin, Chemie, Biologie und Lehre angewendet. Die bekannteste Kategorisierung, die auf der Clusteranalyse basiert, ist das Periodensystem nach Mendelejew, in dem Elemente mit gleicher Eigenschaft zu Gruppen zusammengefasst werden.[9]

Der Vorteil dieses Kategorisierungsverfahrens besteht darin, dass es in vielen Bereichen sowie mit vielen unterschiedlichen Daten angewendet werden kann.[10] Die Clusteranalyse kann mit Hilfe von zwei verschiedenen Verfahren durchgeführt werden. Diese sind das hierarchische und das partitionierende Verfahren.

[8] Vgl. Shmueli (2007, S.1-11)

[9] Vgl. Shmueli (2007, S. 219)

[10] Vgl. Shmueli (2007, S.220)

2.1.2.1 Hierarchische Verfahren

Das **hierarchische Verfahren** kann nochmals in agglomerative und divisive Verfahren unterteilt werden. Beim agglomerativen Verfahren bildet im ersten Schritt jedes einzelne Element eine eigene Gruppe (Cluster). Man erhält also ebensoviele Gruppen wie Elemente. Die beiden Gruppen, die die geringste Distanz zueinander haben, werden zu einer neuen Gruppe (Cluster) zusammengefasst. Diese Gruppenbildung wird solange wiederholt, bis nur noch eine einzige Gruppe (Cluster) vorhanden ist.[11] Man unterscheidet zudem folgende agglomerative Clusteringmethoden:

- Single-Linkage = Minimale Distanz zwischen den Gruppen
- Complete-Linkage = Maximale Distanz zwischen zwei Gruppen
- Average-Linkage = Mittlere Distanz zwischen den Gruppen

Das divisive Verfahren arbeitet im Gegensatz zum agglomerativen Verfahren genau umgekehrt. Alle Elemente bilden eine große Gruppe (Cluster). Diese Gruppe (Cluster) wird nun solange in Gruppen aufgeteilt, bis jede Gruppe nur noch aus einem einzelnen Element besteht.[12]

Das hierarchische Verfahren eignet sich besonders gut, um die Gruppen (Cluster) in eine Hierarchie einzuordnen.[13]

Laut *Shmueli et al.* besitzt das hierarchische Verfahren folgende Nachteile:

- Bei großen Datenmengen wird das Verfahren aufgrund der Berechnungen, die durchzuführen sind, sehr langsam
- Die Datenmenge wird nur einmal durchlaufen, so dass Fehler, die aufgetreten sein könnten, nicht mehr verbessert werden können. [14]

2.1.2.2 Das Partitionierende Verfahren

Partitionierende Verfahren, wie das k-Means Clustering, arbeiten mit einer vorher festgelegten Anzahl an Gruppen. Ziel ist es, eine gegebene Menge in eine vorher festge-

[11] Vgl. Shmueli (2007, S.221f)

[12] Vgl. Shmueli (2007, S.222)

[13] Vgl. Shmueli (2007, S.222)

[14] Vgl. Shmueli (2007, S.232)

legte Anzahl von Gruppen aufzuteilen.[15] Die Ausgangsgruppen bestehen hier ebenfalls aus je einem einzelnen Element. Nun werden die Elemente in jedem Schritt dem nahegelegensten Zentrum einer Gruppe zugeordnet. Dieses Zentrum wird nach der Zuordnung neu berechnet.[16] Bleibt die Zuordnung der Elemente zu den einzelnen Gruppen gleich, wird das Verfahren abgebrochen.[17]

Die Clusteranalyse lässt sich für eine generische Klassifizierung sehr gut einsetzen. Verfahren wie das Data Mining und Text Mining sowie die Taxonomie nutzen dieses Verfahren zur Gruppenbildung.[18]

2.1.3 Taxonomie

Eine weit verbreitete Art, Dokumente und Ressourcen zu kategorisieren, ist die Taxonomie.[19] Der Begriff Taxonomie stammt aus dem Griechischen und lässt sich in die Teile taxis = Ordnung und nomia = Verwaltung zerlegen.[20] Ursprünglich ist die Taxonomie ein Teilgebiet der Biologie, deren Aufgabe es ist, verwandtschaftliche Beziehungen von Tieren und Pflanzen in einem hierarchischen System zu erfassen.[21] Im Wissensmanagement werden Taxonomien häufig als Art Menüstruktur eingesetzt.[22]

[15] Vgl. Shmueli (2007, S.233)

[16] Vgl. Shmueli (2007, S.233)

[17] Vgl. Shmueli (2007, S.233)

[18] Vgl. Priebe (2005, S.1321)

[19] Vgl. Priebe (2005, S.1312)

[20] Vgl. Alby (2007, S.115)

[21] Vgl. Bertelsmann (1994 ,S.143 „Taxonomie“)

[22] Vgl. Priebe (2005, S.1313)

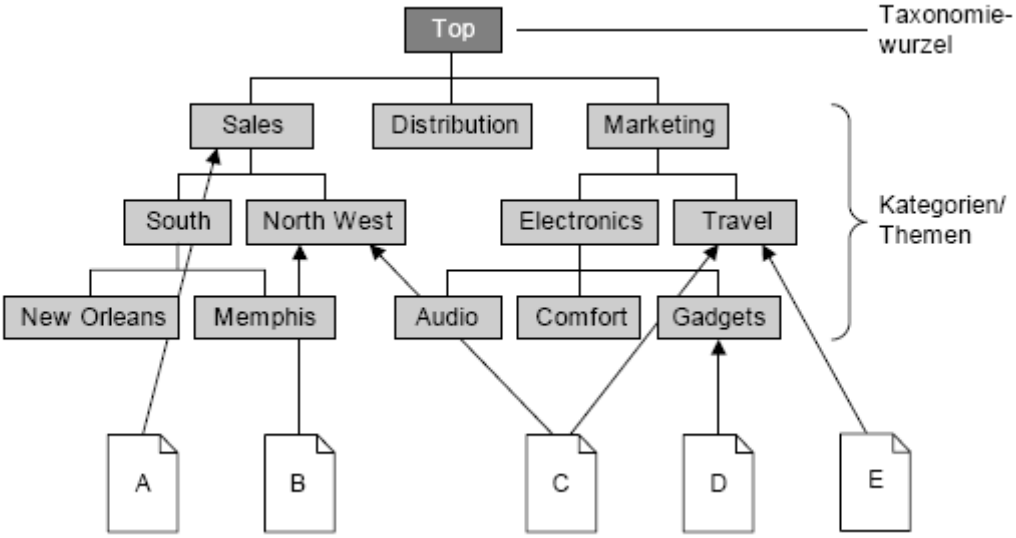


Abbildung 1: Einordnung von Dokumenten in eine Taxonomie [23]

Abbildung 1 zeigt ein Beispiel für Taxonomie in einem Unternehmen, in dem Produkte über ein Call- Center vertrieben werden.[24] Hierbei ist eine Unterteilung in verschiedene Kategorien gegeben. Sales wird hier geografisch weiter in Klassen eingeteilt, wobei Marketing weiter in die Produkte verzweigt.[25] In diese vorgegebene feste hierarchische Struktur können nur Dokumente oder Ressourcen eingefügt werden. Hierbei ist es auch möglich, Dokumente in mehreren Kategorien abzulegen, wie das Dokument C in Abbildung1 zeigt.

Webverzeichnisse wie dmoz[26] basieren auf einer Taxonomie. Dies hat den Vorteil, dass solche Webverzeichnisse jemanden durch ein Thema lotsen können, das vom Umfang her von einer Suchmaschine gar nicht erfasst werden kann wie beispielsweise die Deutsche Außenpolitik.[27] Ein weiterer Vorteil einer Taxonomie ist, dass sich mit ihrer Hilfe Suchergebnisse, wie zum Beispiel die Produktsuche bei eBay, auf bestimmte Kategorien einschränken lassen.[28]

[23] Siehe hierzu Priebe (2005, S.1313)

[24] Vgl. Priebe (2005, S.1312)

[25] Vgl. Priebe (2005, S.1312)

[26] www.dmoz.org

[27] Vgl. Alby (2007, S.120)

[28] Vgl. Alby (2007, S.119)

Nachteil einer Taxonomie ist auf der anderen Seite genau diese vorher festgelegte hierarchische Struktur, da Menschen persönlich alles aus einer „(…)eigenen subjektiven Perspektive“[29] klassifizieren. Wie in Abbildung 1 dargestellt, wird das Unternehmen in drei Bereiche gegliedert. Hierbei wird die geografische Gliederung immer detailierter, desto tiefer man in der Hierarchie nach unten geht. Der Bereich Sales ist nach Nord-Westen und Süden aufgeteilt. Im Süden befinden sich New Orleans und Memphis, „(…) so klassifizieren Menschen alles und ständig, um die einströmenden Informationen verarbeiten und abgelegte Informationen wiederfinden zu können.“[30] In einer Taxonomie ist die Struktur, in der Daten abgelegt werden können, aber schon von vornherein festgelegt, so dass der Anwender sich an diese Struktur anpassen muss.[31]

Nach der Vorstellung der Technik und den Einsatzmöglichkeiten des Verfahrens kann die Taxonomie gut für eine generische Klassifizierung verwendet werden, da sie auch die Fähigkeit besitzt, als Navigationsstruktur zu dienen.

Analysiert man Kategorisierungsverfahren wie die Taxonomie, so stößt man unweigerlich auf die Ontologie, da die Taxonomie nur schwer von der Ontologie zu trennen ist.[32] Eine Taxonomie wird teilweise als einfache Ontologie angesehen. Teilweise wird sie auch als Teilmenge der Ontologie angesehen.[33]

2.1.4 Ontologie in der Informatik

Ursprünglich stammt der Begriff Ontologie aus der Philosophie und „(…)steht dort für die Lehre vom Sein – genauer: von der Möglichkeit und Bedingungen des Seienden-, ist also eng verwandt mit der Erkenntnistheorie, die sich mit den Möglichkeiten und Grenzen menschlichen Wahrnehmens und Erkennens auseinander setzt.“[34]

Ontologie beschreibt einen Wissensbereich …(„mit Hilfe einer standardisierenden Terminologie sowie Beziehungen und ggf. Ableitungsregeln zwischen den dort definierten

[29] Vgl. Alby (2007, S.116)

[30] Siehe hierzu Alby (2007, S.116)

[31] Vgl. Alby (2007, S.118)

[32] Vgl. Priebe (2005, S.1313)

[33] Vgl. Priebe (2005, S.1313)

[34] Siehe hierzu Hesse (2002, S.477)

Begriffen“[35]). Es wird wie in der objektorientierten Programmierung ein generelles Konzept definiert. So kann zum Beispiel die Instanz einer Person, sowie die Beziehung zwischen Personen definiert werden. Wichtige Merkmale einer Ontologie in der Informatik nach *Bodendorf* sind*:*

- Der Rechner ist in der Lage, durch Interpretation der Daten Informationen zu gewinnen.
- Ontologien sind anwendungsunabhängig
- Ontologien enthalten Metawissen und liefern ein Weltbild mit Gegenständen und Naturgesetzen, wobei Gegenstände die Konzepte und die Naturgesetze die Axiome sind.[36]

Außerdem enthalten laut *Bodendorf* alle Ontologien folgende Bestandteile:

- **Konzepte:** Elemente, die über Relationen miteinander verknüpft werden. Um die Elemente ansprechen zu können, werden sie mit einem Symbol versehen.[37]
- **Relationen:** Darstellung von Beziehungen zwischen Konzepten. Es muss zwischen strukturbildenden Relationen und nicht strukturbildenden Relationen unterschieden werden. Strukturbildende Relationen werden als „ist – ein“ Beziehung dargestellt (siehe Abbildung 2). Andere Relationen statten das Konzept mit Eigenschaften wie der Darstellung von Besitzverhältnissen aus.[38]
- **Axiome:** Beinhalten Restriktionen, die durch eine Relation allein nicht dargestellt werden können. Axiome dienen mit „Wenn – dann“ Bedingungen zur Einschränkung von Relationen zwischen Konzepten.[39]

Abbildung 2 stellt einen Ausschnitt aus einer Ontologie dar, die ein Unternehmen beschreibt. Aus der Abbildung ist zu entnehmen, dass Unternehmen Organisationen sind.

[35] Siehe hierzu Hesse (2002, S.477)

[36] Vgl. Bodendorf (2006, S.128)

[37] Vgl. Bodendorf (2006, S.127)

[38] Vgl. Bodendorf (2006, S.127)

[39] Vgl. Bodendorf (2006, S.127)

Personen sind Teil des Unternehmens. Das Unternehmen kauft oder besitzt Wirtschaftsgüter.[40]

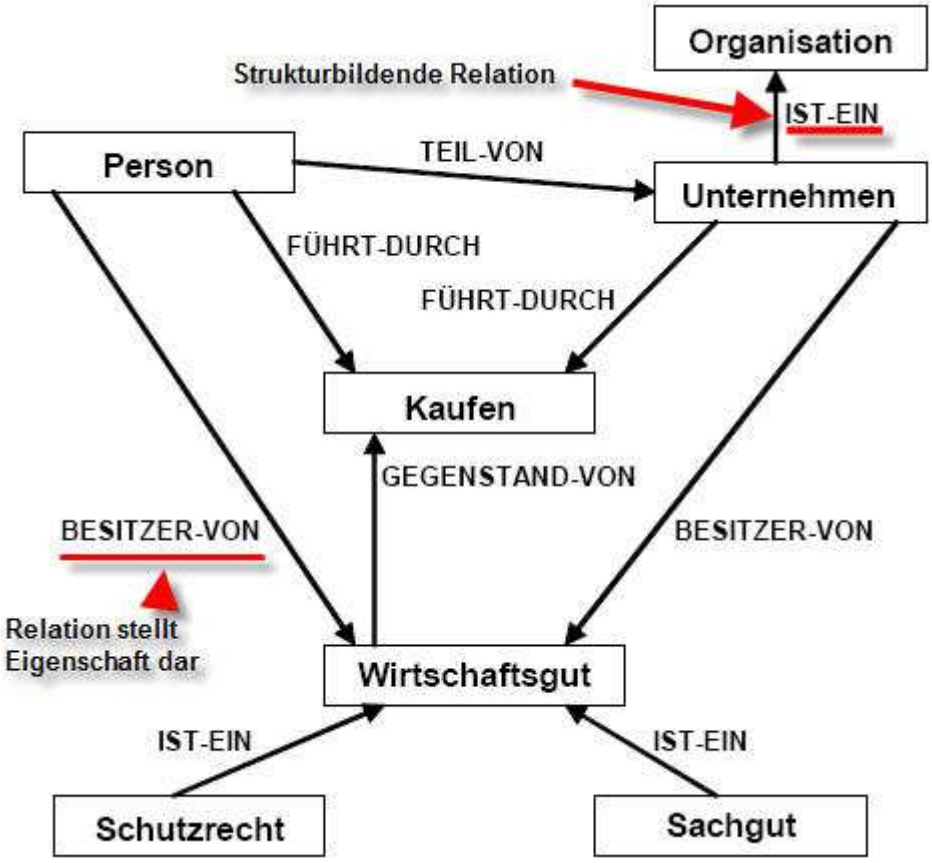


Abbildung 2: Ausschnitt aus einer Ontologie [41]

Ontologien dienen dazu, einen Ausschnitt aus der Realität mit Formalismen, wie Web Ontology Language(OWL) basierend auf RDF (Ressource Description Framework; siehe RDF) zu beschreiben. Die Taxonomie dient in diesem Fall als hierarchische Benutzernavigation, während die Ontologie viel komplexer strukturiert ist und eine höhere Zahl an Instanzen aufweist.[42]

[40] Vgl. Bodendorf (2006, S.126)

[41] Siehe hierzu Bodendorf (2006, S.126) Anmerkungen durch Verfasser hinzugefügt.

[42] Vgl. Priebe (2005, S.1313)

Laut *Steffen Staab* dient die Ontologie im Wissensmanagement dazu, Wissen strukturiert bereitzustellen.[43] Ontologien dienen somit dem Austausch und Teilen von Wissen.[44]

2.1.5 Folksonomie

Nachdem bisher die „klassischen" Klassifikationsverfahren Data Mining, Clusteranalyse, Taxonomie und Ontologie vorgestellt wurden, wird nun das relativ neue Klassifikationsverfahren, Folksonomie betrachtet.

Laut *Alby* handelt es sich bei dem Begriff Folksonomie um einen zusammengesetzten Begriff, bestehend aus dem engl. Begriff: „Folks" (Mensch) und dem Begriff „Taxonomie" (einordnen in eine hierarchische Klassenstruktur)[45] (siehe hierzu auch Kapitel 2.1.3).

In einer Folksonomie werden Ressourcen wie Fotos, Bilder, Bookmarks, Dokumente und vieles mehr durch den einzelnen Anwender selbst klassifiziert. Laut *Schmitz et al.* sind die bekanntesten Beispiele für eine Folksonomie die Webseiten von Flickr[46] und del.icio.us[47]. Diese Systeme sind in ihrem Aufbau ähnlich: Ein Anwender kann sich im jeweiligen System anmelden und Ressourcen in das System einfügen[48]. Diese Ressourcen werden nun mit eigenen Kennzeichnungen (Schlagwörtern) versehen, wobei die hier vergebenen Schlagwörter als „personomy" bezeichnet werden. Alle personomies zusammen bilden die Folksonomie.[49] Im Gegensatz zur Taxonomie muss sich der Anwender nicht an eine vorgegebene hierarchische Struktur anpassen, sondern erzeugt mit der Vergabe von Schlagwörtern eigene Strukturen, die in einer flachen Ebene abgelegt werden.[50, 51]

---

[43] Vgl. Staab (2002, S.194)

[44] Vgl. Staab (2002, S.201)

[45] Vgl. Alby (2007, S.121)

[46] http://flickr.com/

[47] http://del.icio.us/

[48] Vgl. Schmitz (2006, S.262)

[49] Vgl. Schmitz et al. (2006, S.262)

[50] Vgl. Alby (2007, S.121)

[51] Vgl. Mathes (2004, S.4)

In der Folksonomie wird die Vergabe von eigenen Schlagwörtern als „tagging“ bezeichnet, wobei ein einzelnes Schlagwort „tag“ genannt wird.[52] Ein Vorteil der Folksonomie gegenüber der Taxonomie ist, dass sich der Anwender nicht an eine vorgegebene Struktur sowie ein vorgegebenes Vokabular gewöhnen muss, sondern sein eigenes Vokabular verwenden kann.[53] Als Beispiel wird ein Foto vom Pariser Eiffelturm in die Webseite Flickr[54] eingestellt. Der Anwender muss sich nun nicht mehr entscheiden, ob er das Foto in die Kategorie „Paris“, „Eiffelturm“, oder doch lieber „Architektur“ einordnen soll, denn das Foto wurde mit den Begriffen: „Paris, Eiffelturm, Eiffel, Architektur und Stahlbau“ versehen und kann unter diesen Begriffen wiedergefunden werden.[55]

Diese Freiheit führt zu neuen Problemen wie zum Beispiel zur Fragmentierung der Daten. Nehmen wir exemplarisch den Begriff „Blog“, der zum einen im Singular und zum anderen im Plural Verwendung findet. Hieraus entstehen zwei verschiedene Kategorien wie in Abbildung 3 zu sehen ist.[56] Um der Fragmentierung der Daten entgegenzuwirken, zeigt der Anbieter del.icio.us[57] Vorschläge an, welche Schlagwörter andere Benutzer für eine Seite ausgewählt haben.[58]

[52] Vgl. Alby (2007, S.121)

[53] Vgl. Alby (2007, S.122)

[54] http://flickr.com/

[55] Vgl. Alby (2007, S.121-122)

[56] Vgl. Alby (2007, S.122-123)

[57] http://del.icio.us/

[58] Vgl. Alby (2007, S.123)

Abbildung 3: Wortwolke [59]

Die Wortwolke stellt verschiedene Begriffe („Tags“) mit einer unterschiedlichen Schriftgröße für jeden Begriff dar. Desto populärer ein Begriff ist, desto größer wird dieser Begriff dargestellt.[60] Laut *Alby* kann mit Taxonomie eine höhere Präzision beim Wiederfinden und Einschränken von Ressourcen erreicht werden als mit Folksonomie. Ein weiterer Unterschied zwischen der Folksonomie und der Taxonomie ist die Ausgewogenheit der Themengebiete.

Abbildung 4: Webverzeichnis dmoz [61]

[59] http://del.icio.us/tag/

[60] Vgl. Alby (2007, S.122)

[61] http://www.dmoz.org/

In einer Taxonomie sind alle Themen gleichberechtig und auch unpopuläre Themen sind auf der ersten Seite wie bei dmoz[62] vorhanden,[63] siehe Abbildung 4. Zudem werden Taxonomien von Experten gepflegt, so dass auch Teilaspekte von Nischenthemen mit abgedeckt werden.[64] In der Folksonomie dagegen ist eine Ausgeglichenheit der Themen nicht relevant, da nur das in einer Wortwolke erscheint, was populär ist und häufig angeklickt wird.[65] Unpopuläre Themen haben somit wenige Chancen, populärer zu werden.[66]

Nach der Analyse und dem Vergleich der Verfahren Taxonomie und Folksonomie ist der Einsatz der Folksonomie im Bereich der generischen Klassifikation in Kombination mit einer Taxonomie sinnvoll.

2.1.6 Thesauren

Während im Kapitel 2.1.3 die hierarchische Kategorisierung von Ressourcen mit Hilfe von Taxonomien vorgestellt wurde, wird in diesem Kapitel die Kategorisierung von Wörtern, Termen und Begriffen erläutert.

Thesauren definieren ein kontrolliertes Vokabular und stellen Beziehungen zwischen den verschiedenen Termen des Vokabulars her. Sie bilden das sprachliche Gegenstück zu hierarchischen Klassifikationssystemen.[67]

Das kontrollierte Vokabular entsteht durch eine Vielzahl von Zerlegungsschritten.

- **Kontextdefinition:** Zunächst muss der Kontext, den der Thesaurus umfasst, klar abgesteckt werden. Hierzu gehören die Thematik, die Spezifität des Thesaurus, der Sprachstil und der Umfang.
- **Wortsammlung:** Nachdem der Kontext eingegrenzt wurde, müssen aus Quellen Wörter für den Thesaurus entnommen werden. Hierzu eignen sich schon vorhandene Thesauren oder klassifikatorische Systeme, Lehrbücher und andere

[62] http://www.dmoz.org/

[63] Vgl. Alby (2007, S.123)

[64] Vgl. Alby (2007, S.124)

[65] Vgl. Alby (2007, S.124)

[66] Vgl. Alby (2007, S.125)

[67] Vgl. Ferber (2003, S.54 – 58)

Quellen. Die so gesammelten Wörter sollten an dieser Stelle bereits vorklassifiziert werden. Im nächsten Schritt wird das Vokabular einer Synonymkontrolle, Polysemkontrolle und Zerlegungskontrolle unterworfen.

- **Synonymkontrolle:** Bei der Synonymkontrolle werden Synonyme und Schreibweisen- Varianten beseitigt.
- **Polysemkontrolle:** Die Polysemkontrolle überprüft das Vokabular auf Homonyme (Wörter, die identisch klingen oder sich identisch schreiben wie „Lehre“ – „Leere“) und Polyseme (z.B. Schirm für Regenschirm, Bildschirm usw.). Im Thesaurus wird nur ein Bedeutungstil eines Wortes beibehalten.
- **Zerlegungskontrolle:** Die Zerlegungskontrolle überprüft Wörter auf Kompositabbildungen (Donaudampfschifffahrtsgesellschaftskapitän). Es wird je nach Thesaurus entschieden, welche Wörter im Thesaurus zerlegt werden und welche nicht.[68]
- **Äquivalenzklassen:** Die einheitlichen Begriffe, die nach den drei vorhergehenden Kontrollen entstanden sind, werden auch Äquivalenzklassen genannt.
- **Descriptor:** Bei Thesauren mit Vorzugsbenennung wird aus der Äquivalenzklasse ein Begriff, der die gesamte Klasse gut beschreibt, ausgewählt. Dies ist der Descriptor.[69]

Thesauren sind hierarchisch durch eine Oberbegriffs- und Unterbegriffsrelation gegliedert, wobei zu jedem Term ein Oberbegriff und eine Reihe spezifischer Begriffe angegeben werden, wenn sie im Thesaurus existieren. Diese Relationen sind wie eine Klassifikation und können als hierarchischer Graph angesehen werden.

2.1.7 Metadaten

Nachdem Kategorisierungsverfahren wie Data Mining, Clusteranalyse, Taxonomie, Ontologie, Folksonomie und Thesauren betrachtet wurden, werden an dieser Stelle die Metadaten erläutert, ohne die Kategorisierungsverfahren wie die Clusteranalyse und Data Mining nicht funktionieren würden, da diese die Metadaten als Grundlage für die Kategorisierung verwenden.

[68] Vgl. Burkart (2004, S.141ff)

[69] Vgl. Burkart (2004, S.145)

„Metadaten sind Informationen über Dokumente, die mittels eines Feldschemas erfasst werden.“[70] Diese Informationen können dazu genutzt werden, die Suche nach dem Dokument zu verbessern. Aus diesem Grund ist es für digitale Objekte wichtig, dass maschinenlesbare Beschreibungen existieren, damit diese Objekte auch dann gefunden werden können, falls sie nicht schon in einem Katalog einsortiert wurden. Diese Beschreibungen von Objekten werden als Metadaten bezeichnet.[71]

Im folgendem werden die verschiedenen Ansätze und Arten zur Speicherung von Metadaten beschrieben.

2.1.7.1 Dublin – Core – Metadaten

Hierbei handelt es sich um eine Sammlung von Metadatenelementen, die einen Kern von Angaben bilden soll. In Dublin Core ist eine einfache Menge von Elementen definiert. Viele dieser Elemente orientieren sich an bibliographischen Daten für Papierdokumente.

Die Elemente, die das Dublin – Core – Metadatenset enthalten, sind auf der Homepage von Dublin-Core-Metadaten einzusehen. (http://dublincore.org/documents/dces/)

Des Weiteren gibt es folgende Prinzipien beim Auftreten einer Metadatenbeschreibung, die festgelegt wurden:

- **Beschreibung:** Dublin – Core – Elemente beschreiben die Eigenschaft des Objektes selbst
- **Erweiterbarkeit:** Weitere Elemente müssen hinzugefügt werden können, auch wenn das System nicht mit Dublin – Core arbeitet. Diese Elemente müssen aber toleriert werden.
- **Unabhängigkeit von einer spezifischen Syntax:** Es gibt keine festgelegte Syntax für die Angabe der Metadaten.
- **Optimalität:** Es gibt keine Metadatenelemente, die unbedingt angegeben werden müssen.
- **Wiederholbarkeit:** Alle Elemente können mehrfach in einem Datensatz auftauchen, um mehrere Autoren abbilden zu können.

[70] Siehe hierzu Stock (2007, S.402)

[71] Vgl. Ferber (2003, S.267)

- **Veränderbarkeit:** Elemente können durch die Angabe von Attributen verändert werden, wenn eine spezifische Interpretation des Inhaltes gegeben wird. [72]

2.1.7.2 Hierarchisch strukturierte Metadaten

Hierbei handelt es sich um die Learning – Object – Metadata – Spezifikation (LOM Spezifikation). Es ist eine „… hierarchisch strukturierte Beschreibung mit neun Top-Level-Elementen, die sich aus Unterelementen zusammensetzen; also eine Datenstruktur, wie sie durch eine DTD beschrieben wird“[73]. Aufgrund der Datenstruktur mit der Möglichkeit, stark strukturierte Metadatenmodelle zu erstellen, sind komplexe Suchverfahren möglich. Die Datensammlung, die hierfür angelegt und gepflegt werden muss, ist aber sehr aufwändig zu erstellen und zu pflegen.[74]

2.1.7.3 PIC (Platform for Internet Content Selection)

PIC ist einer der ältesten Ansätze, Webdokumente mit Metadaten zu kennzeichnen, und ist der Vorgänger des Resource Description Framework (RDF). Es wurde in erster Linie entwickelt, um Seiten mit jugendgefährdendem Inhalt zu kennzeichnen. Ziel der Kennzeichnung ist es, über eine Negativauswahl Seiten, die zum Beispiel als jugendgefährdend markiert wurden, zu blockieren, und über eine Positivauswahl den Zugriff nur auf nicht jugendgefährdende Seiten oder auf einen Pool von freigegebenen Seiten zu erlauben.[75]

2.1.7.4 RDF (Resource Description Framework)

RDF ist ein Modell, „…das eine Syntax zur Beschreibung der Semantik von Metadaten zur Verfügung stellt,…“.[76] Ziel der Entwicklung von RDF ist es, Metadaten unabhängig von einem Wissensgebiet beschreiben zu können sowie den Austausch von Metadaten

[72] Vgl. Ferber (2003, S.269f)

[73] Siehe hierzu Ferber (2003, S.272)

[74] Vgl. Ferber(2003, S.269f)

[75] Vgl. Ferber (2003, S.276)

[76] Vgl. Ferber (2003, S.276)

zwischen unterschiedlichen Anwendungen zu fördern.[77] RDF basiert auf einem Datenmodell, welches sich aus drei Komponenten zusammensetzt:

- Mit **Ressource oder Objekte (Subjekte)** werden alle „Dinge“ bezeichnet, die man mit RDF beschreiben kann. Diese Dinge können Webseiten, Bücher, Gemälde, Zeitungen usw. sein. Diese Ressourcen werden mit einer URI (Uniform Ressource Identifier) eindeutig bestimmt.[78]
- „Objekte können **Eigenschaften (properties)** haben, die ihre Charakteristika oder Aspekte beschreiben“[79]
- **Aussagen (statements)** werden durch die Verknüpfung von Objekten und Eigenschaften mit der Wertzuweisung einer Eigenschaft erstellt, wobei dieser Wert ebenfalls wieder ein Objekt (im grammatikalischen Sinn) sein kann. Die Aussage kann als Trippel von Subjekt(Ressource), Prädikat(Eigenschaft) und Objekt (grammatikalischer Begriff) bezeichnet werden. Mehrere Aussagen über dasselbe Objekt(Subjekt) werden zu einer Beschreibung(description) zusammengefasst. [80]

Außerdem stellt das Datenmodel noch 3 Container (Bag, Sequence und Alternative) zur Verfügung, in denen Konzepte und Werte zusammengefasst werden können, wobei der Container „Bag“ für ungeordnete Mengen steht, in dem Objekte mehrmals vorkommen können. Der Container „Sequence“ ist für eine geordnete Liste von Objekten. Der Container „Alternative“ stellt eine Menge von Objekten dar, aus der ein Objekt ausgewählt werden muss. [81] Aufgrund des Sachverhaltes, dass Aussagen als Objekte definiert werden können, sind mit dem RDF-Datenmodell auch Aussagen über Aussagen möglich. Dies ist eine Aussage über ein Objekt.[82] RDF wurde mit Hilfe von XML definiert.

[77] Vgl. Ferber (2003, S.276)

[78] Vgl. Ferber (2003, S.277)

[79] Siehe hierzu Ferber (2003, S.277)

[80] Vgl. Ferber (2003, S.277)

[81] Vgl. Ferber (2003, S.277)

[82] Vgl. Ferber (2003, S.277)

2.2 Information Retrieval

2.2.1 Was ist Information Retrieval?

Nach *Womser-Hacker/Mandl* beschreibt Information Retrieval die Suche nach Informationen, die Wiedergewinnung von Wissen aus einer großen Menge von Informationen, die Repräsentation, die Speicherung und die Organisation von Wissen.[83] Bei der Wiedergewinnung von Wissen wird aus dieser Gesamtmenge von Informationen eine Teilmenge ermittelt, die für den Anwender die nötigen Informationen enthält.

Die Datengrundlage für Information Retrieval Systeme bilden dabei jegliche Art von digitalisierten Medien wie Bilder, Dokumente, Texte, Zeitungsartikel usw. Die größte Bedeutung haben an dieser Vielzahl von Medien die Texte. Ein Problem, welches sich bei Texten ergibt ist, dass sie bei Massendaten nicht vollständig erfasst werden können. So wird ein Text auf Terme, Descriptoren und Schlagwörter begrenzt, die keine Beziehung mehr untereinander haben, sodass die Maschine den Zusammenhang, worum es in dem Text geht, nicht mehr herstellen kann. Ist ein Dokument mit den Schlagwörtern Maschine und Herstellung gekennzeichnet, ist der Inhalt des Dokumentes nicht klar ersichtlich. Handelt es sich also um ein Dokument, welches die Herstellung von Maschinen beschreibt, oder handelt es sich um ein Dokument, das den Einsatz von Maschinen in der Herstellung eines Produktes beschreibt?[84]

2.2.2 Ablauf eines Information Retrieval Prozesses

Der Prozess lässt sich als Dialog darstellen. Der Anwender gibt seine Bedürfnisse an das Information Retrieval System weiter, wobei er auf die Möglichkeiten, die das System liefert, beschränkt ist. Die Anfrage wird mit den vorhandenen Dokumenten bzw. Repräsentationen im Information Retrieval System verglichen, wobei das System eine Teilmenge der vorhandenen Daten zurückgibt.

[83] Vgl. Womser-Hacker/Mandl (2007, S.692)

[84] Vgl. Womser-Hacker /Mandl (2007, S.692)

3 Analyse von Kategorisierungen in der Praxis

Im vorherigen Kapitel wurden Begriffe und Verfahren zur Kategorisierung von Ressourcen erläutert, die dem Verständnis in diesem Kapitel dienen. In diesem Kapitel werden auf der Grundlage, der in Kapitel 2 gewonnen Erkenntnisse, Kategorisierungen in einem Dokumenten-Management System auf Basis von IBM Lotus Notes Domino untersucht.

Lotus Notes ist eine Groupware Software, die dem Anwender ein Messaging System, zahlreiche Office-Funktionalitäten sowie ein leistungsfähiges Datenbanksystem zur Verfügung stellt. Es bietet im Datenbanksystem die Möglichkeit, Dokumente auf vielfältige Weise zu kategorisieren. Diese Kategorisierungen werden hierbei in Ansichten (Views) dargestellt und basieren auf Feldern, die in Masken (Forms) angelegt wurden. Das Kategorisierungsverfahren, welches Lotus Notes zugrunde liegt, ist die Taxonomie. Hierbei werden vorhandene hierarchische Strukturen in sortierten Spalten der View angezeigt. Die hierarchische Struktur dient dabei ebenso zur Navigation, wie auch dem Wiederfinden von Daten. Auf den in Lotus Notes vorhandenen Funktionalitäten der Kommunikation und Kategorisierung setzt die Datenbank K-Pool auf. Hierbei handelt es sich um eine Datenbank, die am Lehrstuhl Wirtschaftsinformatik 2 eingesetzt wird und auf dem von der Fima Pavone entwickelten Dokumenten-, Knowledge-Managementsystem „Enterprise Office“[85] basiert. Sie wird zur Organisation, Speicherung und dem Wiederfinden von Wissen eingesetzt. Das Enterprise Office besteht aus folgenden Datenbanken:

- K-Pool Datenbank
- Organisation Datenbank
- Prozess Datenbank
- Settings Datenbank

Für die weitere Analyse werden lediglich die Datenbanken K-Pool und Settings weiter betrachtet. In der Settings Datenbank werden zentrale Einstellungen, Vorlagen sowie Listeneinträge organisiert und gespeichert. Zu diesen Elementen gehören unter anderem

[85] Ausführliche Informationen über Pavone Enterprise Office erhält der geneigte Leser in Nastansky(2002)

die Stichwortlisten[86]. In einer Stichwortliste können mehrere Stichwörter abgelegt werden, wobei Stichwortlisten eine klassenartige Struktur darstellen, da die Werte in der Form „Name der Sichwortliste“ \ „Wert“ gespeichert werden. Die Stichwortliste wird im weiteren Verlauf als Klasse bezeichnet.

Der K-Pool ist die Datenbank, die der Anwender für das gesamte Dokumenten-, Knowledge-Management einsetzt. In dieser Datenbank werden die Dokumente erstellt und gespeichert. Der Fokus dieser Analyse bezieht sich auf die im K-Pool gegebenen Möglichkeiten zur Kategorisierung von Dokumenten.

Die im K-Pool verwendete Kategorisierungs-Technik ist identisch mit der in Lotus Notes vorgestellten Taxonomie (siehe oben und Kapitel 2.1.3) und dient dem Anwender im K-Pool zur Navigation.

Um in Lotus Notes und dem K-Pool eine Kategorisierung zu erstellen, muss eine Spalte in einer Ansicht erstellt werden, die diese Kategorie aufnimmt. Die Spalte muss die Eigenschaft „kategorisiert“ und ein Feld oder eine Formel zugeordnet bekommen, welche die Auswahl der passenden Dokumente ermöglicht. Um Kategorien auf- und zuklappen zu können, muss die Eigenschaft „Show twisties when row is expandable“ ausgewählt sein. Dies führt dazu, dass ein kleines Dreieck vor den Kategorien erscheint.

In einer Ansicht können ebenfalls mehrere kategorisierte Spalten vorhanden sein. Um ein Dokument zu kategorisieren, muss der Name nach dem kategorisiert werden soll in das Feld geschrieben werden. Dies können pro Dokument ein oder mehrere Felder sein.

Um im K-Pool eine möglichst umfassende Übersicht über die vorhandenen Dokumente zu erlangen, sind pro Dokument mehrere Felder für die Aufnahme von Metadaten vorgesehen (vgl. hierzu Kapitel 2.1.7). Im Grundlagen-Kapitel wurde die Taxonomie als eine hierarchische Kategorisierung vorgestellt, in der man sich an eine fest vorgegebene Kategorisierung anpassen muss.

[86] Vgl. Nastansky (2002, S.281)

Im K-Pool existieren folgende Möglichkeiten zur Kategorisierung der Dokumente:

- nach Keywords
- nach Themes
- nach Meta-Structures

Aufgrund der Eigenschaft, dass der Anwender die Möglichkeit hat, seine eigene hierarchische Ordnung in die schon bestehenden Strukturen zu integrieren, handelt es sich hierbei nicht um eine reine taxonomische Kategorisierung.

Diese Kategorisierungsmöglichkeiten sind vornehmlich auf einen Feldtyp beschränkt. Die vorhandenen grafischen Benutzeroberflächen zur Eingabe von Kategorien existieren nur für Keywords und Meta-Structures. Im Anschluss werden die Benutzeroberflächen für die Eingabe der unterschiedlichen Kategorien und deren Eigenschaften genauer betrachtet.

Im K-Pool wurde darauf geachtet, dass Dokumente eine Vielzahl von Metadaten enthalten können, die der Kategorisierung dienen. Hier sind die Felder „NamesReference_input“; „PAVKeywords“ und „PAVCategories“ besonders zu nennen. Dokumente im K-Pool können nach diesen Feldern sortiert in Views dargestellt werden. Um es dem Anwender auf einfache Art und Weise zu ermöglichen, Daten in diese Felder einzufügen, existieren drei verschiedene Benutzeroberflächen. Als erstes wird das Eingeben der sogenannten „Themes“ in das Feld „NamesReference_input“ betrachtet.

Bei „Themes“ handelt es sich um ein Textfeld, mit dessen Hilfe eine Taxonomie aufgebaut wird. Dieses Feld ist für die Eingabe von mehreren Werten ausgelegt. Um eine hierarchische Struktur zum Ablegen des Dokumentes zu erstellen, werden die Begriffe, die eine Hierarchieebene bilden, ausgehend von der höchsten zur niedrigsten Ebene mithilfe eines Backslashs („\“) getrennt.

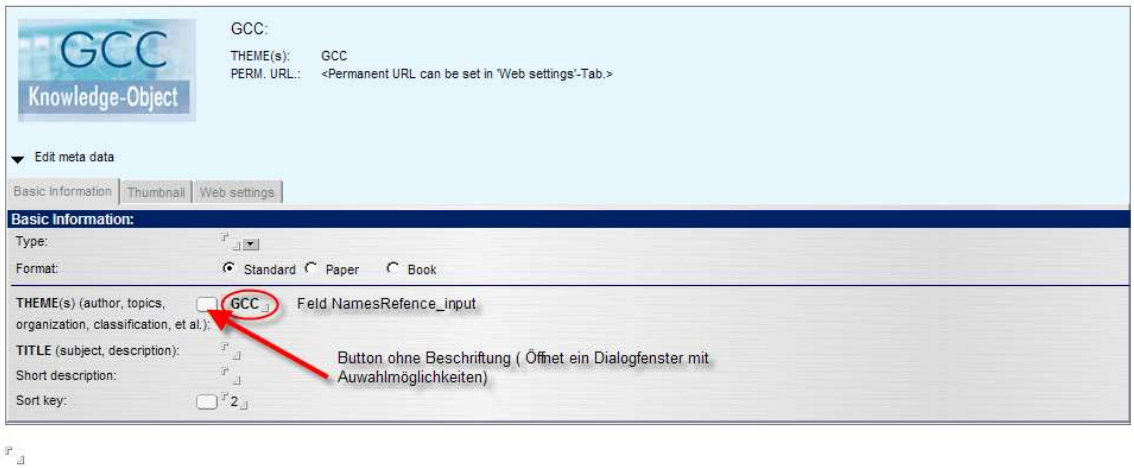


Abbildung 5: Themes; Eingabemöglichkeit [87]

In Abbildung 5 ist die Eingabe von „Themes“ dargestellt. Hierbei können Werte per Hand eingegeben werden. Jedoch ist es für Lotus Notes Domino unerfahrene Anwender an dieser Stelle nicht sofort ersichtlich, wie eine hierarchische Struktur zu erzeugen ist. Es ist ebenfalls nicht zu erkennen, auf welche Weise mehrere Werte einzugeben sind. Mit Hilfe des Buttons, der in Abbildung 5 markiert wurde, hat der Anwender die Möglichkeit, sich eine Auswahlbox einblenden zu lassen. In Abbildung 6 ist eine Auswahlbox mit zwei markierten Einträgen dargestellt. Erkennbar an den beiden mit rot umkreisten Häkchen.

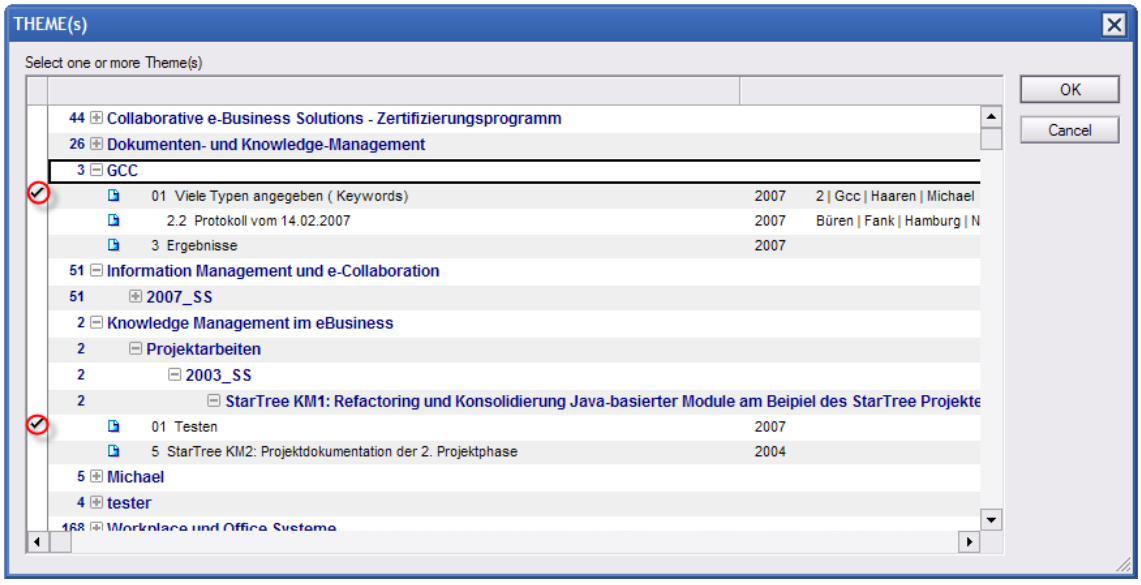


Abbildung 6: Auswahldialog von bestehenden Werten für Feld NamesReference_input [88]

[87] Screenshot aus dem K-Pool

[88] Screenshot aus dem K-Pool

Mit einem Klick auf den OK Button werden die markierten Werte als Themes übernommen. Die Werte sind in diesem Fall: „GCC“ und „Knowledge Management im eBusiness\Projektarbeiten\2003_SS\StarTree KM1:…“ Die Möglichkeit, aus diesem Dialogfenster eine neue Hierarchieebene in das Dokument einzufügen, gibt es hier nicht. Eine solche Ebene muss im Textfeld (siehe Abbildung 5) mit einem „\“ an der entsprechenden Stelle eingefügt werden. Zudem gibt es derzeit keine Möglichkeit, die Kategorien, die in diesem Fall als „Themes“ angelegt wurden, für mehrere Dokumente gleichzeitig zu ändern. Dies bedeutet, dass jedes Dokument einzeln geöffnet werden muss, um die Kategorie anzupassen. Werden mehrere Dokumente angelegt, die alle in derselben Kategorie wie zum Beispiel der Kategorie „GCC“ abgelegt werden sollen, so muss diese Kategorie bei jedem neuen Dokument erneut aus der View ausgewählt, geöffnet und von Hand eingetragen werden.

Als nächstes werden die Kategorisierung nach Keywords und die entsprechende grafische Benutzeroberfläche analysiert. Keywords liegen nicht als normale Schlagwörter vor, sondern sind in Klassen strukturiert und werden in der Settings Datenbank als Stichwortlisten – wie oben beschrieben – angelegt und im K-Pool zur Verfügung gestellt. In der Settings Datenbank können Stichwortlisten mit zwei Eigenschaften erstellt werden. Stichwortlisten können zum einen zusätzliche Begriffe, die vorher nicht definiert wurden zulassen, und zum anderen können die Stichwortlisten auf die vorgegebene Begriffe beschränkt werden. In diesem Fall bilden die Stichwortlisten ein festes Vokabular (siehe Kapitel 2.1.6). Diese Strukturierung soll dem Anwender ermöglichen, die gewünschten Schlagwörter in der Liste zu finden und dem Dokument zu zuordnen. Desweiteren wird der Anwender durch ein Dialogfenster beim Hinzufügen von Keywords (siehe Abbildung 7) unterstützt.

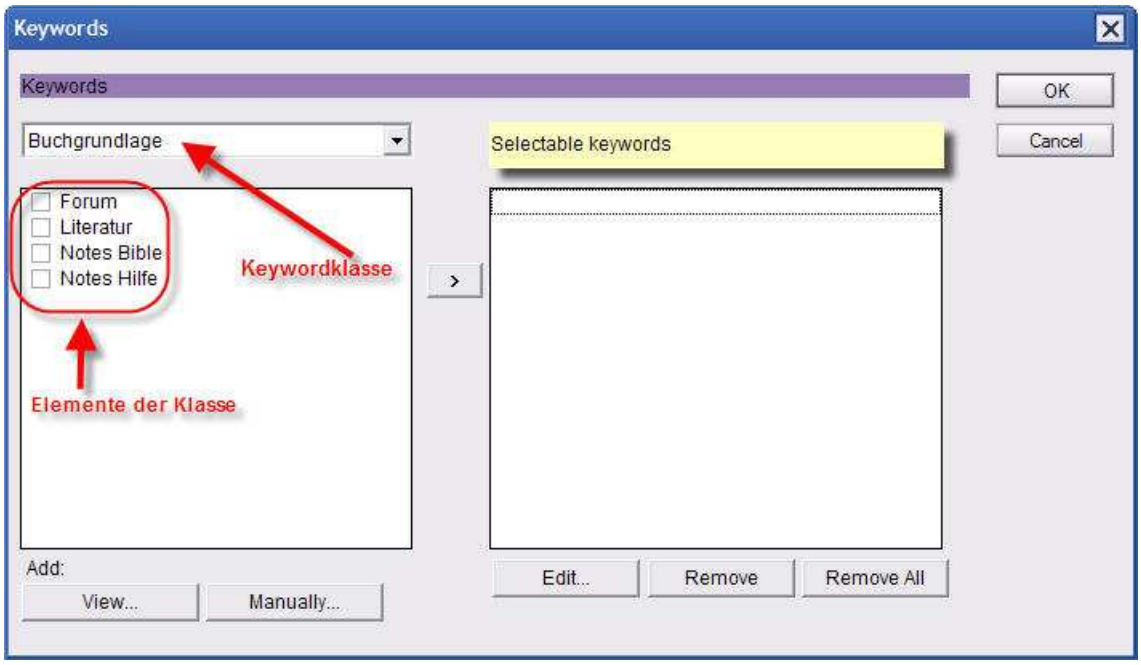


Abbildung 7: Keyword-Auswahldialog [89]

Dieses Dialogfenster (siehe Abbildung 7) ist zudem recht intuitiv zu bedienen. Bei längerer Benutzung fallen einem Poweruser jedoch folgende Schwachstellen des Dialogfensters auf, welche die Arbeit erschweren:

- Eine Gesamtübersicht über vergebene Keywords ist nicht vorhanden.
- Hinzufügen von nur wenigen Keywords aus verschiedenen Klassen gestaltet sich zeitaufwendig und verursacht eine hohe Anzahl von Mausklicks.
- Hinzufügen von Keywords, die noch nicht in der Vorgabeliste enthalten sind, ist zeit- und arbeitsaufwendig, da hierzu jeweils der Button „Manually“ betätigt werden muss. In dem Fenster, welches sich an dieser Stelle öffnet, gibt es jeweils nur die Möglichkeit, ein einziges neues Keyword einzugeben.
- Stehen mehrere vorgegebene Elemente zur Verfügung, so muss der User das gewünschte Element aus den vorhandenen Elementen heraussuchen. Dies ist bei einer Anzahl von über 30 Elementen sehr zeitaufwendig.
- Ausgewählte Elemente bleiben nach Übernahme in das Dokument weiterhin markiert. Dies hat zur Konsequenz, dass die Elemente von Hand deselektiert werden müssen, um weiterarbeiten zu können.

[89] Screenshot K-Pool Keyword-Auswahldialog

Das Hinzufügen von Keywords entspricht einer Mischung aus Taxonomie und einer abgewandelten Form einer Folksonomie. Die Keywords sind in einer zweistufigen Hierarchie organisiert (Klasse \ Wert). Zu den Klassen können einzelne Werte hinzugefügt werden. Um eine Fragmentierung der Begriffe zu vermeiden, werden schon vorhandene Begriffe angezeigt (siehe 2.1.5).

Im nächsten Abschnitt werden die Meta-Structures analysiert. Sie dienen ebenfalls zum Anlegen einer hierarchischen Struktur. Dem Anwender wird zum Anlegen von Kategorien ein Dialogfenster (siehe Abbildung 8) zur Unterstützung des Vorgangs angezeigt, in dem auch bereits vergebene Kategorien ersichtlich sind. Im Dialogfeld sind alle Möglichkeiten zur Texteingabe durch Buttons dargestellt.

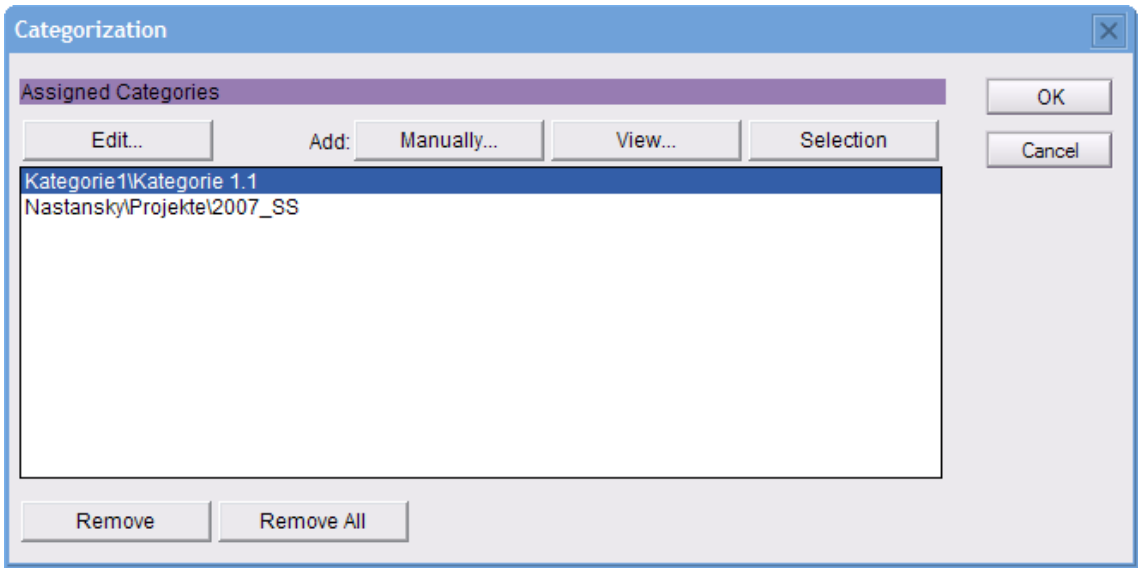


Abbildung 8: Meta-Structures Dialogfenster [90]

Auch das Anlegen einer hierarchischen Struktur wird durch den Dialog unterstützt, indem der Anwender für jede Hierarchieebene eine neue Zeile nutzt. Das in Abbildung 9 gezeigte Beispiel legt eine Hierarchie mit folgenden Ebenen an:

Universität
→ Wirtschaftsinformatik
→ Sommersemester 2007

Durch diese Eingabemaske können die Werte ebenfalls durch andere Zeichenketten anstatt der in Lotus Notes Domino üblichen „\“(Backslash) Zeichen getrennt werden, ohne dass sich der Endanwender über diesen Punkt Gedanken machen muss.

[90] Screenshot K-Pool

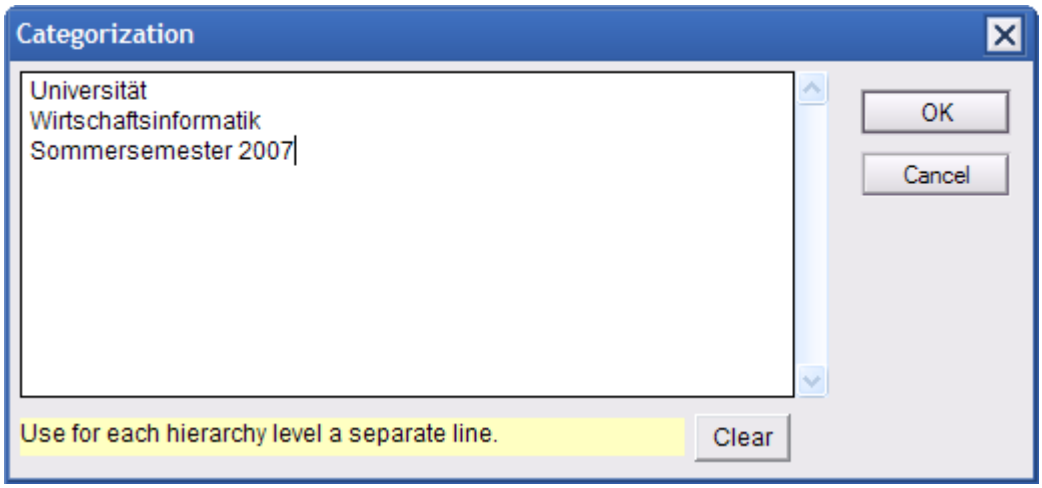


Abbildung 9: Eingabe von Hierarchieebenen

Bei den oben genannten Kategorisierungen handelt es sich um Textfelder, die eine Liste von Feldwerten enthalten können. Die grafischen Benutzeroberflächen sind an diese Felder angepasst und können somit leider nicht für andere Datenbanken verwendet werden. Das Editieren und Hinzufügen von Keywords sowie Meta-Structures ist im K-Pool bereits für mehrere Dokumente aus einer Ansicht heraus implementiert. In einer View werden die zu bearbeitenden Dokumente ausgewählt und mit einem Button der Keyword-Dialog oder Meta-Structures-Dialog geöffnet. Das Besondere des aus der View gestarteten Dialogfensters ist, dass Keywords oder Meta-Structures, die nicht in allen ausgewählten Dokumenten enthalten sind, mit zwei eckigen Klammern vor und hinter dem Keyword oder Meta-Structures hinzugefügt werden. Dies dient der Übersichtlichkeit und gibt dem Anwender die Information, dass er Dokumente ausgewählt hat, die sich in verschiedenen Kategorien befinden.

3.1 Anforderungskatalog

Aus den Analysen in Kapitel 3 und der Anforderung der Generik lässt sich folgender Anforderungskatalog für ein generisches Klassifizierungswerkzeug erstellen.

Das System soll folgenden Anforderungen genügen:

- Das Kategorisierungs-System soll mit sämtlichen Listenfeldern zusammenarbeiten.
- Das System soll eine einfache, leicht zu bedienende Benutzeroberfläche bieten. Zudem soll die Benutzeroberfläche das Hinzufügen von Keywords / Kategorien übersichtlicher gestalten und dabei möglichst wenig Mausklicks erfordern.
- Das Kategorisierungs-System soll sowohl in einer Form als auch in einer View Kategorisierungen durchführen können.
- Es soll ein Unterschied zwischen dem Lese- und Bearbeitungsmodus dargestellt werden.
- Das neue Klassifizierungssystem soll in jeder Lotus Notes Datenbank einsetzbar sein.

4 Vorstellung des Konzeptes

Nachdem in den vorherigen Kapiteln die Grundlagen für Kategorisierungs-Systeme vorgestellt wurden und das System analysiert worden ist, wird nun das Konzept für ein generisches Kategorisierungs–System vorgestellt. Die Analyse in Kapitel 3 stellte heraus, dass zwei verschiedene Arten der Kategorisierung –mit einer Klassenstruktur und ohne Klassenstruktur– vorhanden sind. Aufgrund dieser Tatsache wurden für jedes System Konzepte erstellt, die in den folgenden Kapiteln erläutert werden. Zudem dient der in Kapitel 3.1 erstellte Anforderungskatalog als Grundlage für das Kategorisierungs-System. Auf dieser Basis und der in Kapitel 3 angestellten Analyse der Schwachstellen wurden die folgenden Kategorisierungs-Systeme konzipiert.

4.1 Kategorisierung ohne Klassenstruktur

In Kapitel 3 wurde festgestellt, dass eine grafische Oberfläche im K-Pool, die dem Anwender viele Möglichkeiten für die Bearbeitung von Kategorien zur Verfügung stellt, bereits implementiert ist. Dieses Konzept dient als Ausgangsbasis für die Entwicklung eines generischen Kategorisierungs-Systems. Um dem Anwender die Suche nach Funktionen wie zum Beispiel: Hinzufügen, Ändern, Editieren und Löschen von Kategorien zu erleichtern, sollten diese auf einer Oberfläche dargestellt werden.

Abbildung 10 stellt ein Konzept für eine grafische Benutzeroberfläche dar, das zur Unterstützung der Kategorisierung dient. Bei diesem Konzept wird an der Taxonomie zugunsten der Navigation festgehalten (vgl. 2.1.3). Die Oberfläche soll die Möglichkeit bieten, die bereits einem Dokument zugewiesenen Kategorien darzustellen. Desweiteren müssen bestehende Kategorien editierbar sein und neue Kategorien müssen hinzugefügt werden können. Dabei soll dem Anwender – zur Vereinfachung –eine manuelle Eingabe von Kategorien, sowie eine Auswahl aus schon vorhandenen Kategorien ermöglicht werden. Es ist davon auszugehen, dass der Anwender mehrere Dokumente nacheinander einer Kategorie zuordnen wird. Daher werden die zuletzt verwendeten Kategorien in einem separaten Bereich dargestellt. Dies bringt mit sich, dass eine Kategorisierung von Dokumenten schneller durchführbar ist.

.

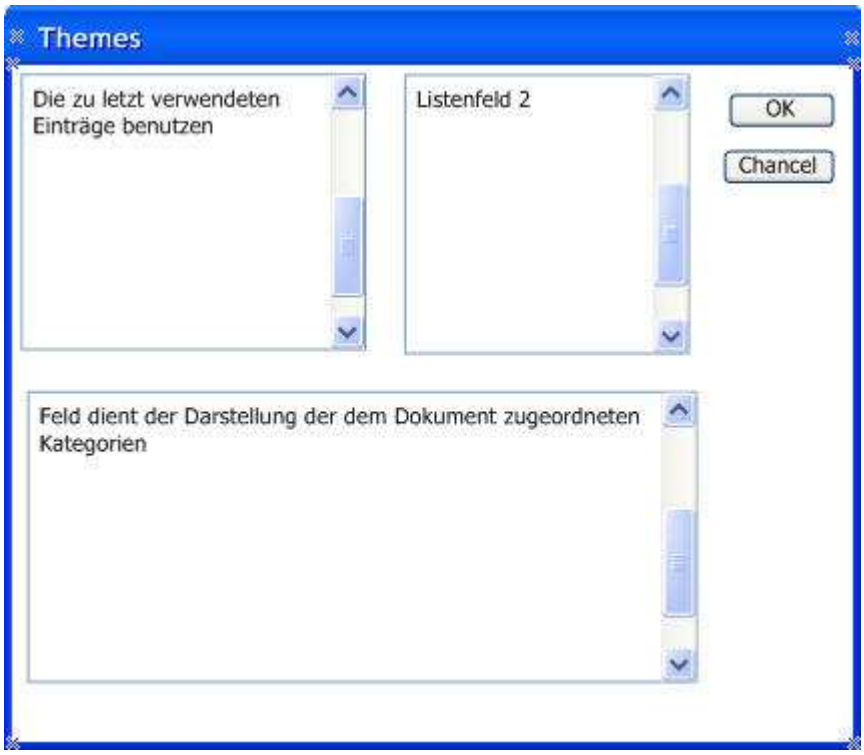


Abbildung 10: Oberflächenkonzept für Kategorien

Um die Übersichtlichkeit bei einer Mehrfachauswahl von Dokumenten zu erhöhen, soll ein separates Feld zur Verfügung stehen, welches Kategorien anzeigt, die nicht in allen Dokumenten vorhanden sind. Zusätzlich soll dies im unteren Feld von Abbildung 10 durch das Einschließen der Kategorie in „[[ ]]“ (wie schon in der Ausgangsbasis dargestellt) angezeigt werden.

4.2 Kategorisierung mit Klassenstruktur

Nach der Analyse des Keyword-Dialoges sowie der Analyse der Kategorisierung muss ein Konzept entworfen werden, welches versucht, allen Anforderungen aus dem Kapitel 3.1 gerecht zu werden. Um dies zu erreichen, ist eine grafische Benutzeroberfläche erforderlich, die eine umfassende Übersicht über existierende Klassen sowie den in den Klassen vorhandenen Werten gibt. Hierzu wurde zunächst folgendes Konzept entwickelt, welches anhand von Abbildung 11 erläutert wird.

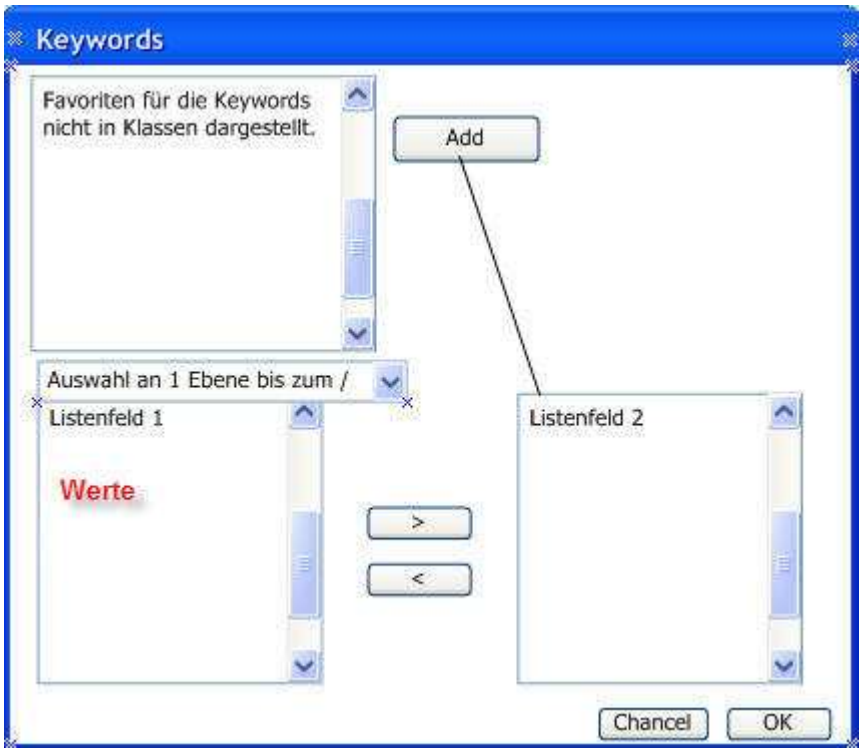


Abbildung 11: Keyword Konzept 1

Abbildung 11 zeigt den ersten Konzeptentwurf. Dieser Entwurf nutzt die im vorrangegangenen Konzept vorgestellte Idee, die zuletzt genutzten Kategorien wieder auf einer Oberfläche zur Verfügung zu stellen. Dieses Konzept ist dem im Kapitel 3 vorgestelltem System für Keywords sehr ähnlich, aber erfüllte nicht den Anspruch der Übersichtlichkeit. Aus diesem Grund wurde ein weiteres Konzept entwickelt.

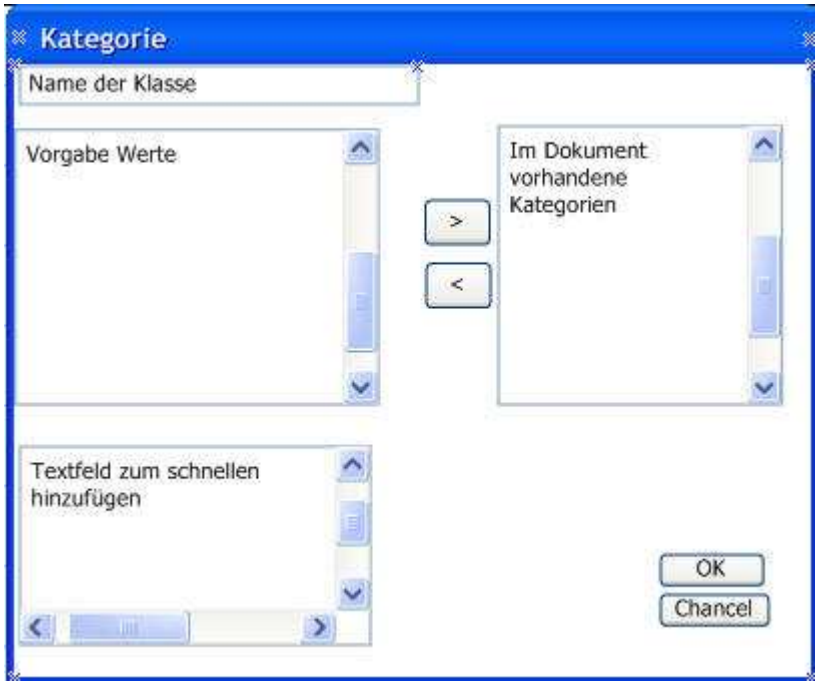


Abbildung 12: Keyword Konzept 2

Abbildung 12 zeigt ein anderes Konzept für die Unterstützung der Kategorisierung. Dieses stellt eine Art Modul dar, welches für jede Klasse implementiert werden soll. So entsteht eine Übersicht über alle vorhandenen Werte in der jeweiligen Klasse und den schon vergebenen Kategorien, die in den Dokumenten enthalten sind. Zudem soll ein Feld, wie es im unteren Teil von Abbildung 12 dargestellt ist, mehrere Werte aufnehmen können, die auf Knopfdruck dem Dokument in Verbindung mit der Klasse hinzugefügt werden. Diese Module sollen dann wie folgt in einer Tabelle angeordnet werden: jeweils zwei Module nebeneinander und mehrere Module untereinander. Durch diese Anordnung wird die geforderte Übersicht über die Kategorien erzielt. Hierbei soll die Anzahl der anzuzeigenden Module auf die Anzahl der vorhandenen Klassen beschränkt werden. Daraus folgt, dass die Klassen, Werte und die Anzahl der Module, die angezeigt werden, jedesmal erneut berechnet werden müssen. Zudem müssen die Vorgaben aus der Stichwortliste der Settings Datenbank mit in dieser Auswahl angezeigt werden. Um bei einer Mehrfachkategorisierung aus einer View die Unterschiede in den Kategorisierungen der Dokumente darstellen zu können, ist es bei diesem Konzept nicht sinnvoll, ein separates Feld einzuführen. Stattdessen wird die Idee, Kategorien in Klammern einzuschließen, übernommen. Aus der Analyse in Kapitel 3 wissen wir, dass eine Mehrfachauswahl für Dokumente bereits implementiert ist. Auf diese Technik greift das neue Konzept nun zurück.

5 Prototypische Implementierung

In den vorherigen Kapiteln wurden eine Analyse und die Konzepte eines generischen Kategorisierungs-Systems vorgestellt. In diesem Kapitel wird das Ergebnis der Implementierung dargestellt. Besonders sollen in diesem Kapitel die generischen Bestandteile des Systems noch einmal verdeutlicht werden.

Zunächst wird auf die Neuerungen in der View eingegangen, und im Anschluss werden beide Kategorisierungs-Systeme ausführlich erläutert. In der View (siehe Abbildung 13), wurden zwei neue Buttons eingefügt, mit deren Hilfe die Kategorisierung von Dokumenten vorgenommen wird. Dieselben Buttons sind in die Form zum Anlegen von Dokumenten eingefügt worden und starten das jeweils identische Kategorisierungs-System.

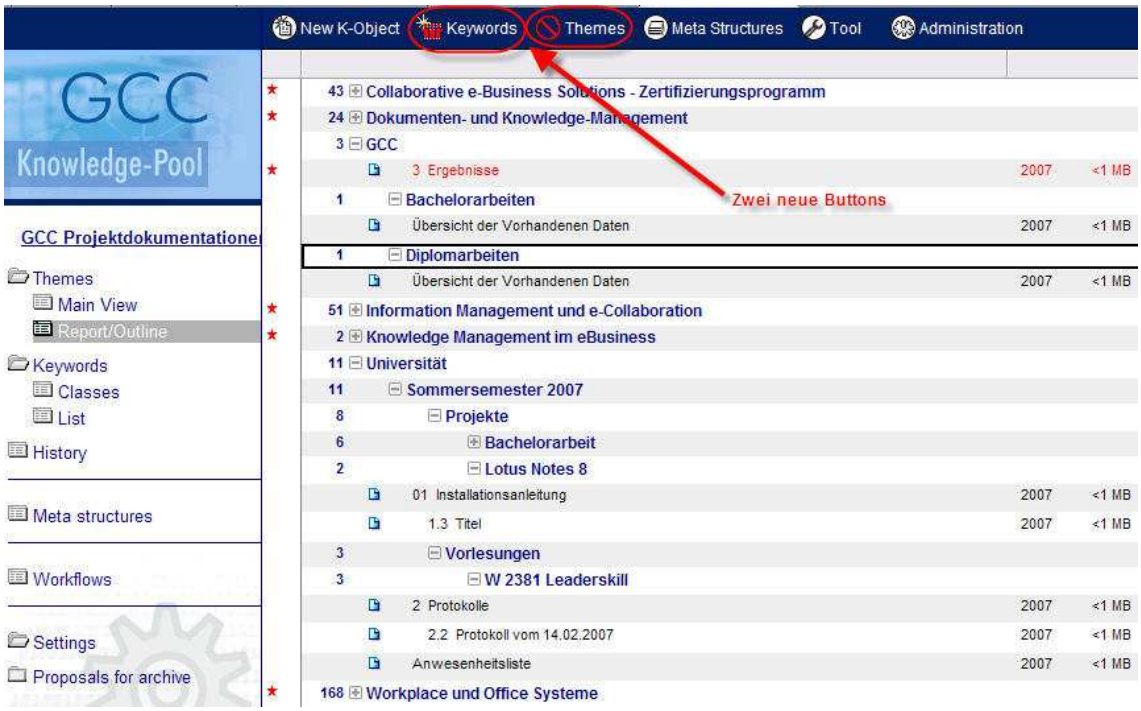


Abbildung 13: View mit neuen Buttons

Das Kategorisierungs-System, das in Abbildung 13 über den Button „Themes“ gestartet wird, basiert auf dem in Kapitel 4.1 vorgestellten Konzept. Werden in einer View mehrere Dokumente aus verschiedenen Kategorien ausgewählt, deren Kategorisierung angepasst werden soll, so wird der Dialog wie folgt dargestellt (siehe Abbildung 14).

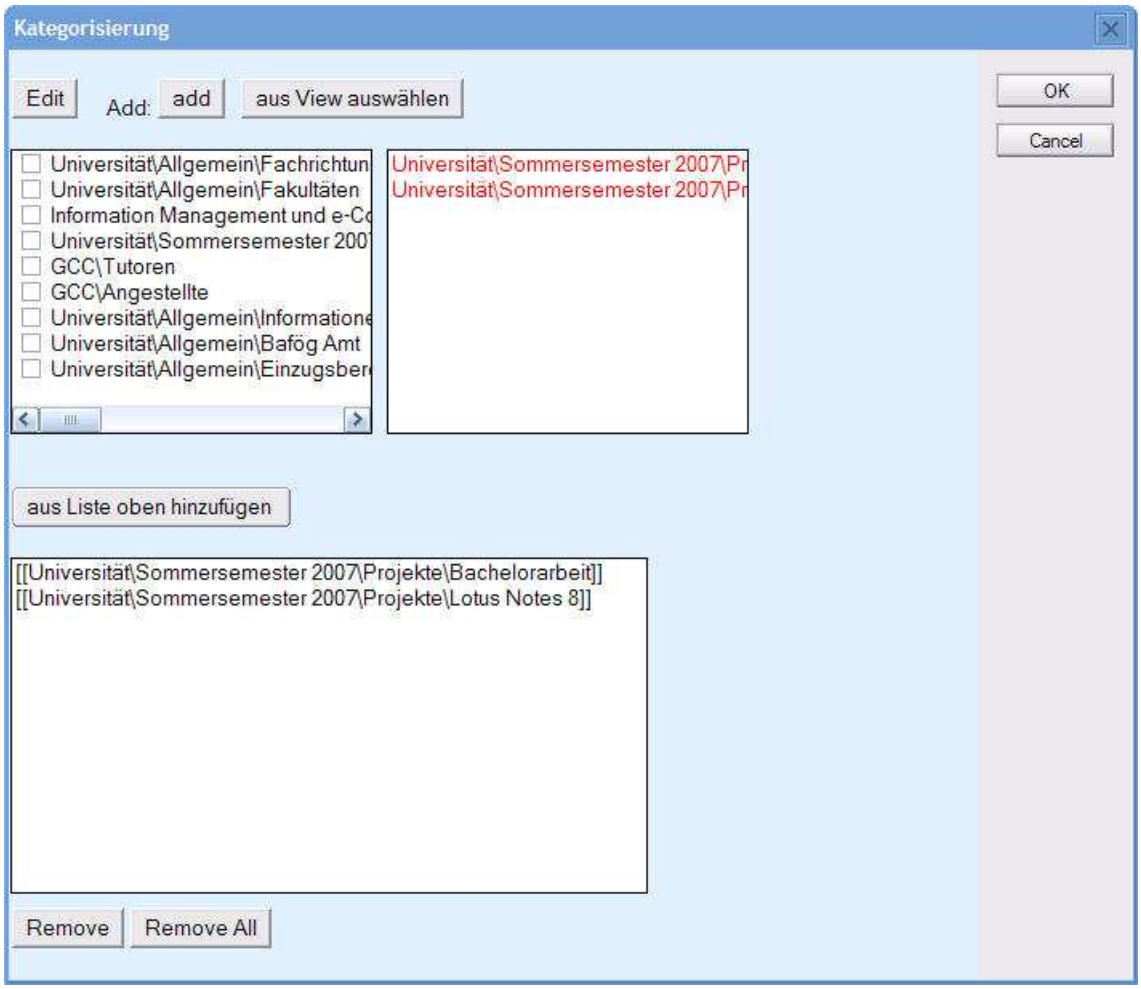


Abbildung 14: Themes Dialog Kap5

Abbildung 14 zeigt das implementierte generische Kategorisierungs-System. Für dieses Beispiel wurden zwei Dokumente aus verschiedenen Kategorien in der View ausgewählt. Im unteren Feld sind die eckigen Klammern zu erkennen, die signalisieren, dass die Dokumente in unterschiedlichen Kategorien eingeordnet sind. Im oberen rechten Feld werden diese Kategorien erneut angezeigt, damit der Anwender genau sieht, dass die Dokumente aus unterschiedlichen Kategorien stammen. Mit dem Edit Button kann die im unteren Feld ausgewählte Kategorie bearbeitet werden. Hierbei öffnet sich ein Fenster, das die Kategorie anzeigt. In diesem Fenster wird für jede Hierarchieebene eine neue Zeile verwendet. Dies erhöht die Übersichtlichkeit, ist aber schon aus dem bisherigen System bekannt. Wird der Button „add“ gedrückt, wird dasselbe Fenster wie in Abbildung 15 geöffnet. Kategorien, die über die View und den „add“ Button hinzugefügt werden, werden in einem Profildokument, das für jeden Anwender angelegt wird, abgespeichert. Hierbei handelt es sich um ein Array, dessen Größe vom Designer bestimmt wird. Diese Kategorien werden dem Anwender beim nächsten Start des Kategorisie-

rungs-Systems im Feld oben Links angeboten, sodass auf schon vergebene Kategorien schnell zugegriffen werden kann.

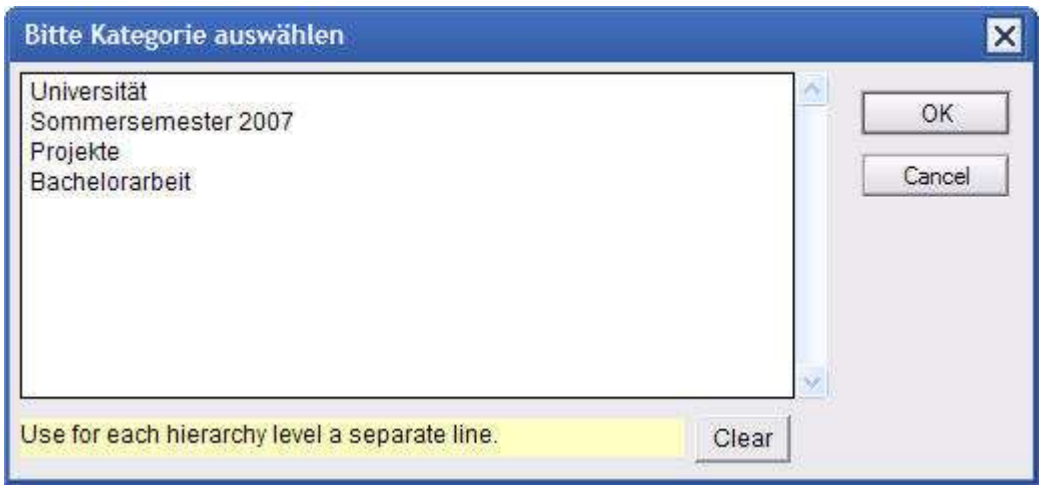


Abbildung 15: Themes Dialog Kap5 Eingabefenster

Um dieses Kategorisierungs-System in vielen Datenbanken einsetzen zu können, wurden alle Funktionen der Buttons in Lotus Script geschrieben. Dies bietet den deutlichen Vorteil, dass Konstanten wie Feldnamen an einer zentralen Stelle in der Form verwaltet werden können.

Im folgenden Abschnitt wird die prototypische Implementierung des Kategorisierungs-Systems mit Klassenstruktur erläutert. Dieses Kategorisierungs-System wird über den in der View befindlichen Button „Keywords“ gestartet. Um die gesamte Funktionsweise des Dialoges darzustellen, wird zunächst auf die im Konzept erläuterte Modulstruktur eingegangen (siehe Abbildung 16).

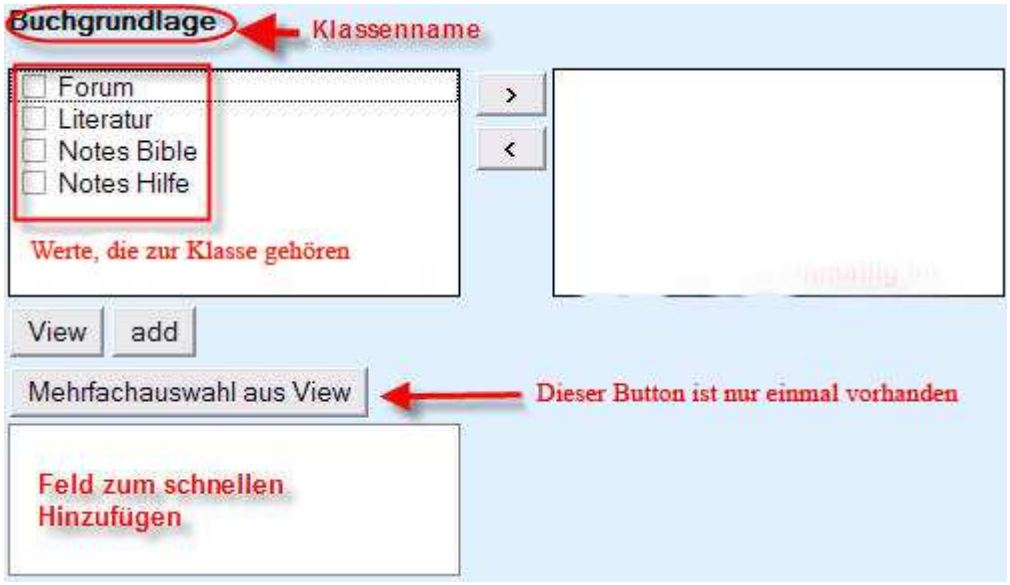


Abbildung 16: Modul des Klassendialoges

In Abbildung 16 ist ein Modul aus dem Kategorisierungs-System mit Klassenstruktur abgebildet. Der Name der Klasse wird hierbei automatisch beim Öffnen der Form berechnet. Der Klassenname dient nun der weiteren Berechnung der Werte, die zur Klasse gehören. In diesem Fall ist der berechnete Name der Klasse „Buchgrundlage“, zu dem die Werte „Forum“, „Literatur“, „Notes Bible“ und „Notes Hilfe“ zählen. Diese Werte dienen dem Anwender als Vorgabewerte, aus welchen der Anwender nun seine Auswahl treffen und über den Button „>“ direkt dem Dokument zuordnen kann. Diese Vorgabewerte wirken dem Problem der Defragmentierung, wie in der Folksonomie beschrieben, entgegen. Bei dieser Implementierung handelt es sich sowohl um eine Taxonomie als auch um den Ansatz einer Folksonomie. Begründet daraus, dass der Anwender zu den einzelnen Klassen eigene Werte hinzufügen kann. Das Hinzufügen von eigenen Werten ist deutlich vereinfacht worden. Die Werte, die zur Klasse und dem Dokument hinzugefügt werden sollen, können in das untere Feld in Abbildung 16 eingetragen werden. Hierbei muss jeder Eintrag in einer einzelnen Zeile eingefügt werden. Zur Eingabe in das Dokument werden die hier eingetragenen Werte über den „add“ Button dem Dokument hinzugefügt. In diesem Dialogfenster ist zusätzlich ein Button vorhanden, der die Beschriftung „View“ trägt. Dieser Button dient dazu, einzelne Werte aus einer hierarchisch sortieren Ansicht auszuwählen und dem Dokument in der entsprechenden Klasse zuzuordnen. Zudem ist in Abbildung 16 ein Button mit der Beschriftung „Mehrfachauswahl aus View“ vorhanden. Dieser Button existiert in der gesamten Form nur einmal, da er dem Anwender die Möglichkeit gibt, mehrere Kategorien direkt aus einer hierarchisch sortierten Ansicht auszuwählen. Diese Auswahl muss in einem zweiten Schritt erneut bestätigt werden, da bei der Auswahl, aufgrund der Architektur von Lotus Notes, Werte nicht eindeutig aus dem Dokument gelesen werden können. Nachdem nun die Funktionsweise des Moduls erläutert wurde, wird die Funktionsweise des gesamten Dialoges erklärt.

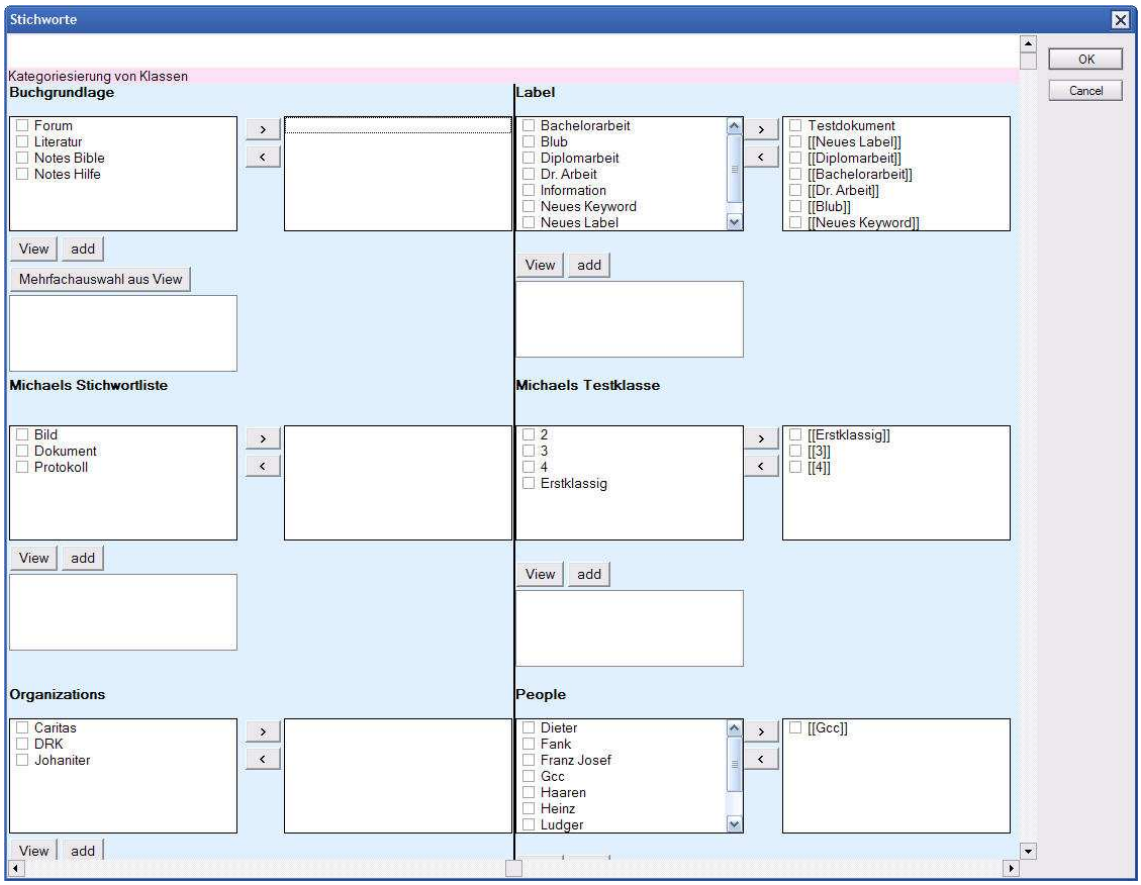


Abbildung 17: Klassenstruktur Dialog

Dieses Kategorisierungs-System schafft durch die einzelnen Module eine Übersicht, wie sie in der vorherigen Version nicht vorhanden war (vgl. Kapitel 3). Die Anzahl der dargestellten Klassen und deren Bezeichnungen werden automatisch berechnet. Durch die Namen der Klassen können die Werte der Klassen berechnet werden. Die Werte zu jeder Klasse werden generisch über den Klassennamen ausgelesen und zu den entsprechenden Feldern unterhalb der Klassenbezeichnung zugeordnet. *Bsp.: Der Name der Klasse ist in Abbildung 16 „Buchgrundlage“. Dieser Name wird automatisch berechnet und dem entsprechenden Feld als Wert übergeben. Mit dem Namen der Klasse ist es nun möglich, die zugehörigen Klassenwerte aus einer View auszulesen und zur Auswahl in einem Feld zur Verfügung zu stellen.*

Dieses Dialogsystem ist derzeit für 20 verschiedene Klassen, die generisch berechnet werden, ausgelegt. Sollten mehr als 20 Klassen in der Datenbank vorhanden sein, so müssen weitere Module eingefügt und Nummerierungen in den Designelementen angepasst werden.

Um das Kategorisierungs-System so zu gestalten, dass es in mehreren Datenbanken und Feldern Verwendung finden kann, wurde die Sprache Lotus Script verwendet, da in

dieser Sprache die Möglichkeit existiert, globale Variablen für ein Designelement zu definieren. Desweiteren wurden Variablen, die in der Formelsprache benötigt werden, in separaten Feldern in der Form abgelegt, sodass diese ebenfalls an einer zentralen Stelle angepasst werden können.

Dieses System baut auf den schon in der Datenbank eingesetzten Verfahren wie im Konzept erläutert auf. Der bereits vorhandene Code, wie zum Beispiel der Aufruf der Funktionen aus der View heraus, wurde leicht für dieses System angepasst. Für die weiteren Funktionen wurde der bestehende Code als Ausgangsbasis verwendet. Zudem mussten sämtliche Buttons, die in der Ausgangssituation in @Functions beschrieben wurden, auf Lotus Script angepasst werden.

Angabe der Codestellen:

Der Quellcode dieser Arbeit ist in der Datenbank K-Pool auf der beigelegten CD in den Forms „1 KategorieUI“ und „1 Kategorie Klassenstruktur GTP20“ sowie in der View „GCC_NC_Themes compact“ in den Buttons „Keywords“ und „Themes“ zu finden. Die globalen Variablen wurden im Globals-Bereich der Designelemente abgelegt. Zur besseren Übersicht über den gesamt verwendeten Quellcode wurden zwei Libraries angelegt (Themes und Keywords) die den Code nochmals enthalten. Im Options-Bereich der Libraries befinden sich kurze Informationen zu den einzelnen Codeelementen.

Probleme:

- Die Anforderung, dass ein Unterschied zwischen dem Editiermodus und Lesemodus des Dokumentes bei Verwendung des Kategorisierungs-Systems dargestellt wird, konnte beim Kategorisierungs-System mit Klassenstruktur leider nur unzureichend implementiert werden. Der Fehler konnte leider bisher nicht entdeckt werden.
- Leider ist der Einsatz des Systems noch nicht mit Nummern- und Datumsfeldern möglich.
- Aufgrund von nicht geklärten Umständen wird das Postopenevent beim Öffnen eines Dokumentes aus der Form heraus nicht ausgeführt. Um das Problem zu kompensieren, wurden zwei Buttons in die Form implementiert, die dieselbe Funktion wie das Postopenevent erfüllen. Drücken Sie hierzu erst den unbeschrifteten Button und anschließend den „Refresh“ Button.
- Eine Fehlermeldung beim öffnen des Keyword-Dialoges im Lese-Modus.

6 Ausblick

Diese beiden Kategorisierungs-Systeme schöpfen den gesamten Umfang der Kategorisierungen noch nicht aus, die in Kapitel 2.1 recherchiert wurden. Zunächst muss die Akzeptanz des Systems von den Usern ermittelt werden. Setzt sich dieses System durch, so können auf Basis der vergebenen Keywords zum Beispiel Wortwolken genutzt werden, um die häufigsten Themen darzustellen. Mit Hilfe der Keywords könnte zudem ein leistungsstarkes Suchverfahren implementiert werden, welches das Wiederfinden von Dokumenten unterstützt. Diesen Information Retrieval Ansatz zu implementieren, würde nach geleisteter Vorarbeit im Bereich der Kategorisierung umfangreich bis sehr umfangreich ausfallen. Ein Kategorisierungs-System ist jedoch der erste Schritt zum Information Retrieval.

Da das Kategorisieren jedoch immer noch eine gewisse Zeit benötigt, könnte auf Basis von Data (Text) Mining Verfahren eine automatische Verschlagwortung von noch nicht kategorisierten Dokumenten eingeführt werden.

7 Fazit

Mit der Implementierung der Kategorisierungs-Systeme wird den zukünftigen Anwendern eine Übersicht geboten, die sie in dem vorherigen System nicht hatten. Durch diese Übersicht ist ein schnelleres sowie gezielteres Arbeiten möglich. Ob das System deutliche Vorteile gegenüber den bisherigen Möglichkeiten bietet, wird sich erst im Laufe der Zeit herausstellen. Wie die Erfahrung zeigt, werden Schwachstellen erst nach längerem Einsatz einer Anwendung deutlich.

Mit diesem System wird sich vermutlich auch die Moral der Anwender erhöhen, Dokumente mit allen wichtigen Daten zur Kategorisierung zu füllen. Diese Annahme beruht auf der Tatsache, dass der Anwender alle möglichen Klassen und alle möglichen Optionen, die durch Buttons dargestellt sind, direkt präsentiert bekommt, um sie zu nutzen.

Desweitern liefert die Einführung einer Darstellung der zuletzt benutzten Kategorisierungen eine erhebliche Zeitersparnis, wenn viele Dokumente angelegt werden.

Da durch diese Implementierung nun ein direktes Ändern von Kategorien aus einer View heraus erfolgen kann, können viele Dokumente einfacher geändert werden. Dies bedeutet eine hohe „Klick-“ und Zeitersparnis für den Anwender.

Literaturverzeichnis

[Alby 2007]

Alby, Tom: Web 2.0 – Konzepte, Anwendungen, Technologien, 2. aktualisierte Ausgabe, Hanser Verlag, München, 2007

[Benz / Oliver 2003]

Benz, Brian; Rocky Oliver: Lotus Notes and Domino 6 Programming Bible, Wiley Publishing, Inc. Verlag Indianapolis, Indiana 2003

[Bertelsmann 1994]

Lexikon-Institut Bertelsmann: Bertelsmann Lexikon Band 14 Stick-Venn, Bertelsmann Lexikothek Verlag GmbH, Gütersloh 1994, S.143 „Taxonomie“

[Bodendorf 2006]

Bodendorf, Freimut: Daten- und Wissensmanagement, Zweite, aktualisierte und erweiterte Auflage, Springer Verlag, Berlin Heidelberg, 2006

[Ferber 2003]

Ferber, Reginald: Information Retrieval – Suchmodelle und Data- Mining- Verfahren für Textsammlungen und das Web, dpunkt.verlag, Heidelberg 2003

[Gruber 1993]

Gruber, Thomas R: A Translation Approach to Portable Ontology Specifications – Technical Report KSL 92-71

KNOWLEDGE SYSTEMS LABORATORY
Computer Science Department
Stanford University
Stanford, California 94305

[Hesse 2002]

Hesse, Wolfgang: Ontologie(n), In: Informatik Spektrum, Heft Volume 25, Number 6 / Dezember 2002, Springer Verlag Berlin / Heidelberg 2002

[Nastansky 2002]

Nastansky, Ludwig; Bruse, Thomas; Haberstock, Philipp; Huth, Carsten; Smolnik, Stefan: Büroinformations- und Kommunikationssysteme: Groupware, Workflow Management, Organisationsmodellierung und Messaging-Systeme. In: Fischer, Joachim; He-

rold, Werner; Dangelmaier, Wilhelm; Nastansky, Ludwig; Suhl, Leena: Bausteine der Wirtschaftsinformatik - Grundlagen, Anwendungen, PC-Praxis, 3., überarbeitete Auflage, Erich Schmidt Verlag, Berlin 2002, S. 235-324

[Mathes 2004]

Mathes, Adam (2004): Folksonomies - Cooperative Classification and Communication Through Shared Metadata
http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.pdf Zugriff am 16.08.2007

[Priebe 2002]

Priebe, Torsten; Kolter, Jan; Kiss, Christine: Semiautomatische Annotation von Textdokumenten mit semantischen Metadaten, Wirtschaftsinformatik 2005, Physica-Verlag, Heidelberg 2005, S.1309 - 1328

[Schmitz 2006]

Schmitz, Christoph; Hotho, Andreas; Jäschke, Robert; Stumme, Gerd (2006): Mining Association Rules in Folksonomies, In: Data Science and Classification. Hrsg.: Batagelj, Vladimir; Bock, Hans-Hermann; Ferligoj, Anuska; Ziberna, Ales, Springer-Verlag Berlin Heidelberg 2006, S.261-270

[Shmueli 2007]

Shmueli, Galit; Patel, Nitin R.; Bruce, Peter C.: Datamining for Business Intelligence; John Wiley & Sons, Inc Verlag, Hoboken, New Jersey 2007

[Staab 2002]

Staab, Steffen: Wissensmanagement mit Ontologien und Metadaten, In: Informatik Spektrum, Heft Volume 25, Number 3 / Juni 2002, Springer Verlag Berlin / Heidelberg 2002

[Stock 2007]

Stock, Wolfgang: Information Retrieval – Informationen suchen und finden; Oldenburg Wissenschaftsverlag GmbH, München 2007

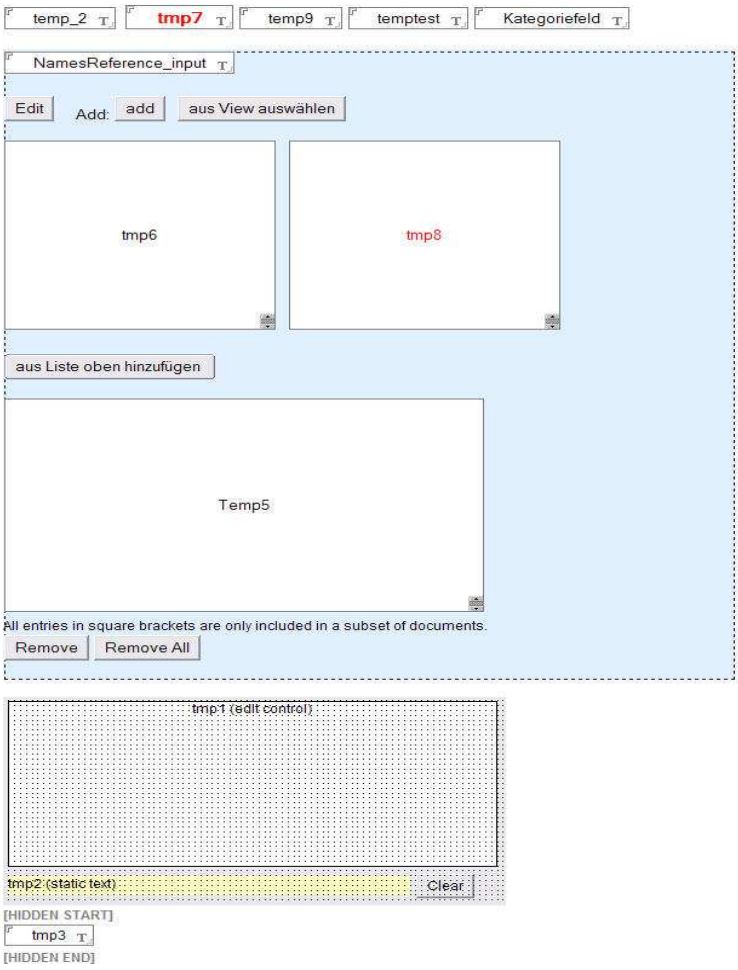
[Womser-Hacker /Mandl 2007]

Womser-Hacker, Christa; Mandl, Thomas: Information Retrieval; Das Wirtschaftsstudium (WISU) Ausgabe 5/07, Lange Verlag, Düsseldorf 2007, S. 692-697

Anhang A

A.1 Installationsanleitung ohne Klassenstruktur

Um das generische Kategorisierungswerkzeug in einer Datenbank zu installieren, müssen folgende Schritte durchgeführt werden. Die Installation wird am Beispiel des K-Pool erläutert.

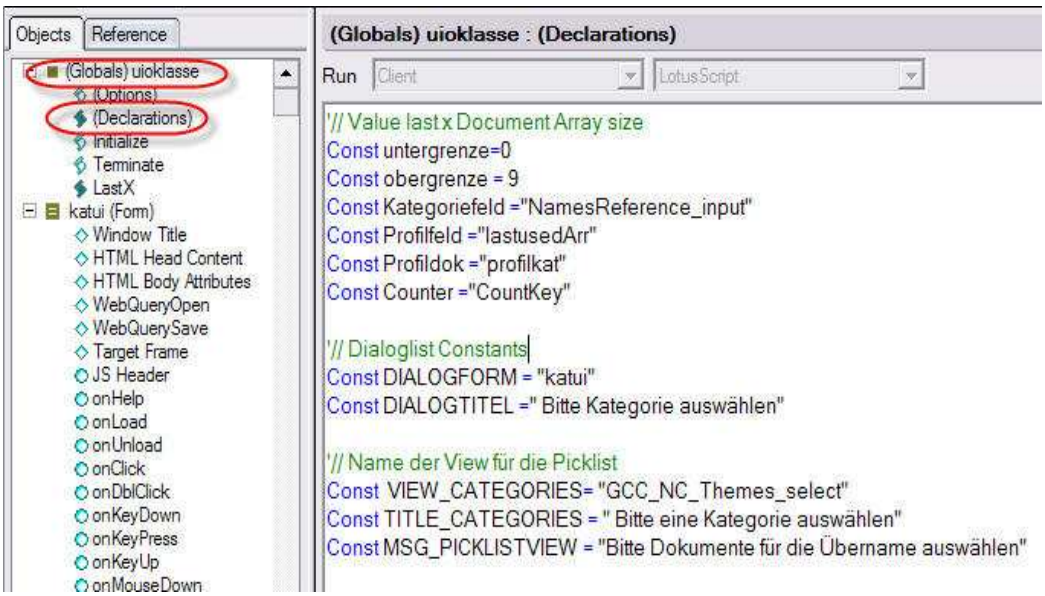
1. Öffnen Sie die Lotus Notes Datenbank, die das zu verwendene Kategorisierungs-System enthält und die Datenbank, in der Sie das Kategorisierungs-System einfügen wollen, im Lotus Domino Designer.
2. Kopieren Sie die Maske (Form) „1 KategorieUI“ mit dem Alias „katui“ in die Datenbank und öffnen Sie die Form. Sie sollten nun folgendes Bild sehen.
3. 

4. In dieser Form müssen einige Einstellungen vorgenommen werden. Da im K-Pool Themes auf Grundlage des Feldes „NamesReference_input“ vergeben werden, ist es erforderlich, dass Sie das Feld, wie in der Abbildung oben zu sehen

ist, mit dem entsprechenden Namen des Feldes, das ihrer Kategorisierung dient zu benennen und den Namen des Feldes als Defaultwert eingeben. Dieser Feldname muss ebenfalls im Feld „Kategoriefeld“ eingegeben werden.

5. Die Felder „temp_2“ und „tmp7“ enthalten Informationen über Profildokumente. Sollten Sie noch keine zweite Variante dieser grafischen Oberfläche eingefügt haben, so müssen diese Felder nicht angepasst werden. Andernfalls müssen Sie hier einen anderen Namen für das Profilfeld eingeben.

6. Anpassungen der globalen Variablen der Form werden im „Globals“ Bereich durchgeführt



7. Um die globalen Variablen einstellen zu können, scrollen Sie –wie im Bild oben dargestellt– im Reiter Objects ganz nach oben, bis sie dort den Eintrag Globals lesen. Hier können Sie – wie im Bild oben zu sehen ist – den Namen des Kategoriefeldes sowie die Größe des Arrays mit den Einträgen „untergrenze“ und „obergrenze“, der zuletzt verwendeten Kategorien einstellen. Hierbei wird empfohlen, nur den Wert für die Obergrenze zu verändern. Zudem können Sie ebenfalls den Namen des Profilfeldes, des Profildokumentes und die Ansicht (View), die als Grundlage für eine Mehrfachauswahl genutzt werden soll (hier „GCC_NC_Themes_select“), eingeben.

8. Im Feld „tmp3“, welches Sie unten auf der Form sehen, wird der Rückgabewert speziell für den K-Pool modifiziert. Sollten Sie keine Besonderheiten bei der Rückgabe benötigen, wie es zum Beispiel in der JournalDatenbank ist, so kom-

mentieren Sie den derzeit aktiven Code aus und nutzen die im Feld angegebene Alternative.

9. Kopieren sie nun den Button „Themes new“ in die entsprechende View.
10. Als nächstes kopieren Sie die Konstanten aus der Standard-View, die unter Globals\Declarations vorhanden sind, ebenfalls in die Standard-View ihrer Datenbank und geben Sie hier die Namen der entsprechenden Felder ein.
11. Öffnen Sie im Anschluss die View, sodass Sie das neue Kategorisierungs-System nutzen können.

A.2 Installationsanleitung mit Klassenstruktur

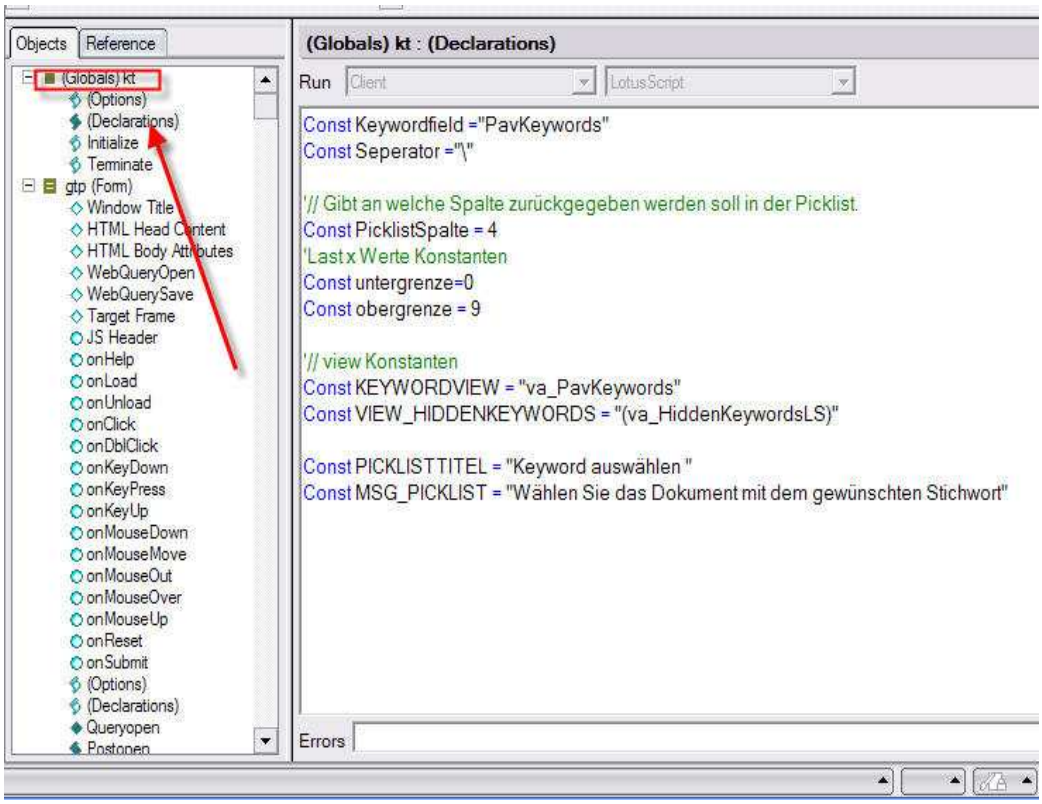
An dieser Stelle wird die Installation des generischen Kategorisierungs-Systems mit Klassenstruktur am Beispiel des K-Pools erläutert. Hierfür sind folgende Schritte durchzuführen:

1. Öffnen Sie die Lotus Notes Datenbank, die das zu verwendene Kategorisierungs-System enthält und die Datenbank, in der Sie das Kategorisierungs-System einfügen wollen, im Lotus Domino Designer.
2. Kopieren Sie die Form „1 Kategorie Klassenstruktur GTP20“ mit dem Alias „gtp“ in ihre Datenbank und öffnen Sie die Form.
3. In dieser Form sind nun unterschiedliche Einstellungen vorzunehmen.

4. Die Felder, die in der Abbildung oben rot markiert sind, müssen ausgefüllt werden, da sie der Konfiguration dienen und in der gesamten Form bekannt sind.

5. Das Feld oben links trägt den Namen „PAVkeywords“ und enthält den Namen der Kategorien. Ändern Sie den Namen entsprechend ihres Feldes, das sie für die Kategorisierung verwenden. Im K-Pool kann der Name wie gegeben beibehalten werden.
6. In das Feld „KategorieFeldName“ muss der Name des Feldes eingetragen werden, welches zur Kategorisierung verwendet wird. Der Name der Kategorie muss hierbei in Anführungszeichen eingegeben werden.
7. Das Feld „KeywordViewName“ muss den Namen der View enthalten, die die Kategorien in folgender Form darstellen:

8. Sollte so eine View noch nicht in Ihrer Datenbank vorhanden sein, legen Sie bitte eine View an, die in der ersten Spalte den Namen der Klasse und in der zweiten Spalte die Werte der zugehörigen Klasse enthält. In einer dritten Spalte geben Sie bitte den Namen des Feldes an, welches die Kategorien enthält.
9. Im Feld „Seperator“ wird das Trennzeichen zwischen dem Klassennamen und dem Wert angegeben. Dies ist in diesem Fall ein „\“
10. In das Feld „DbKeywordView“ muss der Name der View eingetragen werden, die die Kategorie nach dem in Punkt 8 beschriebenem Schema enthält.
11. In das Feld „SettingKeywordView“ muss der Name der View eingetragen werden, die die Kategorien in der Settings Datenbank nach diesem Schema liefert. Sollten Sie alle Kategorie in derselben Datenbank angelegt haben, so können sie dieses Feld mit einem Leerstring versehen.
12. Im Feld ColumnDBNr, muss die Nummer der Spalte eingetragen werden, die die Werte der Kategorien enthält. Haben Sie die View so angelegt wie unter Punkt 8 beschrieben, sollte diese Zahl eine zwei sein.
13. Geben Sie in das Feld SettingsCloumnNr die Spalte ein, in der sich die Werte der Klassen befinden.
14. In das Feld KlassenColumnNr steht der Wert 1, wenn sie die View wie unter Punkt 8 beschrieben angelegt haben.

15. 

16. In der Abbildung sind die Globalen Lotus Script Einstellungen für die Form zu sehen. An dieser Stelle müssen Sie den Namen des Feldes eingeben, welches zur Kategorisierung verwendet werden soll. In der oberen Abbildung ist dies PavKeyword. Es muss zudem der Name der View eingegeben werden, der die Kategorien –wie in Punkt 8 beschrieben– enthält.

17. Kopieren Sie den Button „Keyword new“ aus der Standard-View des K-Pools in die Standard-View Ihrer Datenbank.

18. Kopieren Sie die Konstanten aus der Standard-View, die unter Globals\Declarations vorhanden sind, ebenfalls in die Standard-View Ihrer Datenbank. Danach geben Sie die identischen Werte, die Sie in den Globalen Einstellungen der Form eingetragen haben, an.

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides Statt, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Paderborn, den ....

(Datum) (Unterschrift)