nova-net Werkstattreihe

Innovation in der Internetökonomie

Technologie-Roadmap

Mihály Jakob
Fabian Kaiser
Holger Schwarz

Stuttgart 2006

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Herausgeber: Mihály Jakob, Fabian Kaiser
Holger Schwarz
Verlag: Fraunhofer IRB Verlag
Nobelstaße 12, 70569 Stuttgart

ISBN: 3-8167-7047-9

Erscheinungsjahr: 2006

Auslieferung und Vertrieb: Fraunhofer IRB Verlag
Nobelstraße 12
70569 Stuttgart
Telefon +49 (0) 711/9 70-25 00
Telefax +49 (0) 711/9 70-25 08
www.irb.buch.de
www.publica.fhg.de

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Ausgangssituation und Zielsetzung

Wissen, Informationen und Daten haben in den letzten Jahren enorm an Bedeutung gewonnen und spielen mittlerweile in allen Lebensbereichen eine zentrale Rolle. Sie werden mit einer noch nie da gewesenen Geschwindigkeit erzeugt, verbreitet und wieder verwendet. Nicht nur die Datenmenge wächst tagtäglich, auch die entsprechende Infrastruktur zur Informationsverbreitung wird verstärkt ausgebaut. Satteliten, Breitbandnetze sowie Mobiltelefone, die Ton, Bilder und Video übermitteln, speichern und anzeigen, sind nur einige Beispiele. Hierbei spielt das Internet als Massenspeicher und Datenautobahn eine zentrale Rolle. Diese Entwicklung wird durch die Entstehung von Begriffen, wie *Informations- und Wissensgesellschaft* begleitet.

Unternehmen müssen sich an diese geänderte Situation anpassen. Einerseits bieten neue Informationsquellen im Internet großes Potenzial, andererseits drohen wichtige Informationen in der Masse unterzugehen. Obwohl externe Informationsquellen, vor allem im Bereich des Innovationsmanagements, eine wichtige Rolle spielen, ist die strukturierte Erschließung der unternehmensinternen Wissensbasis als Aufgabe höchster Priorität einzustufen.

Die Entwicklungen der letzten Jahre haben nicht nur eine erhöhte Daten- und Informationsmenge mit sich gebracht, die von Unternehmen bewältigt werden muss; unsere Gesellschaft hat sich auch in eine *Informations- und Wissensgesellschaft* gewandelt. Kunden und Endverbraucher sind besser informiert, und haben ein erhöhtes Anspruchsniveau. Dies hat zur Folge, dass Kunden in immer kleiner werdenden Abständen mit neuen, innovativen Produkten bedient werden müssen. Dies führt wiederum zur Verkürzung von Produktions- und Innovationszyklen.

Unternehmen, die sich diesen neuen Herausforderungen erfolgreich stellen wollen, müssen sowohl externe als auch interne Informationsquellen und Informationsverarbeitungsprozesse professionell handhaben. Nur durch den Einsatz eines Wissensvorsprungs kann der nötige Innovationsvorsprung erarbeitet werden, der das Unternehmen im Auge des Kunden von der breiten Masse abhebt.

Wichtige Aufgaben im Bereich des Innovationsmanagements, die Unternehmen ihren Zielen näher bringen, sind die systematische Beobachtung von Trends und Marktentwicklungen (Trendmonitoring), das Einbinden von Kunden in die Produktentwicklung (Lead User Integration) und die systematische Betrachtung und Steuerung des Produktlebenszyklus (Life Cycle e-Valuation).

Weitere Rahmenbedingungen, die Unternehmen in den letzten Jahren verstärkt beachten müssen, sind die des nachhaltigen Wirtschaftens. Das nachhaltige Wirtschaften kann zudem starke Bezüge zur Innovationsfähigkeit eines Unternehmens aufweisen, wenn es einem Unternehmen gelingt, durch seinen Wissensvorsprung innovative und ressourceneffiziente Produkte zu entwickeln.

Im Rahmen der vorliegenden Technologie-Roadmap werden Basis-, Integrations- und Anwendungstechnologien aufgezeigt, die die nachhaltige Gestaltung von Innovationsprozessen unterstützen können. Kapitel 2 stellt den Innovationsprozess und grundsätzliche Themenfelder, die im Forschungsprojekt *nova-net* bearbeitet werden, vor. In Kapitel 3 wird der so genannte *Information Workflow* vorgestellt, der den Informationsverarbeitungsprozess in sinnvolle Phasen unterteilt und die systematische Lösung von Innovationsproblemen ermöglicht. In Kapitel 4 werden für das Innovationsmanagement relevante Technologien den einzelnen Informationsverarbeitungsphasen zugeordnet, während Kapitel 5 sich speziellen Problemen widmet, die in den Themenfeldern des Forschungsprojekts *nova-net* besondere Beachtung verdienen.

2 Der Innovationsprozess

Innovationsmanagement umfasst nach Hauschildt dispositive Tätigkeiten zur Gestaltung einzelner Innovationsprozesse (vgl. Hauschildt 1997). Diese Tätigkeiten werden im Rahmen von Innovationsprojekten ausgeführt. Die wichtigsten Aufgaben sind hierbei:

- Strategien und Ziele definieren und verfolgen,
- Entscheidungen treffen,
- Informationsflüsse bestimmen und beeinflussen,
- soziale Beziehungen herstellen und gestalten.

Ein Innovationsprojekt ist die Umsetzung eines Innovationsvorhabens, das die erfolgreiche Vermarktung einer Invention (Erfindung) zum Ziel hat. Dabei können sich Inventionen auf Produkte, Prozesse oder Organisationen beziehen (vgl. Spath 2004). Den Innovationsprozess kann je nach Betrachtungsweise in drei oder vier Phasen unterteilt werden (vgl. Abbildung 2-1). Das gröbere Raster verwendet die folgenden drei Phasen:

- **Strategisches Management**

  In der strategischen Orientierungsphase muss das Ziel des Innovationsprojektes festgelegt werden. Dafür ist die gründliche Analyse sowohl externer als auch interner Informationsquellen notwendig. Die Ergebnisse dieser Analysen dienen als Entscheidungsgrundlage für die Wahl der Zielsetzung.

- **Ideenmanagement**

  Die Gewinnung von Innovationsideen sowie deren Bewertung und Auswahl stellen die drei zentralen Aufgaben im Ideenmanagement dar. Die in der Orientierungsphase festgesetzte Zielrichtung dient hierbei als Maßstab.

- **Projektmanagement**

  In der Phase Projektmanagement werden schließlich diejenigen Ideen realisiert, die in der vorangegangenen Phase ausgewählt wurden. Wichtige Phasenabschnitte bilden die Projektvorbereitung, Projektplanung und Projektrealisierung. Als Querschnittstätigkeit soll das Projektcontrolling das Projektmanagement unterstützen.

Im Forschungsprojekt *nova-net* haben sich während der Zusammenarbeit mit ausgewählten Praxispartnern drei Schwerpunktfelder ergeben, die vertieft untersucht werden. Diese Schwerpunktfelder und ihre Positionierung im Innovationsprozess werden in Abbildung 2-1 dargestellt und in den folgenden Abschnitten kurz erläu-

tert. Detaillierte Erläuterungen finden sich in Arbeitspapieren des Projekts *nova-net* (vgl. Fichter und Kiehne 2004, Lang et al. 2004 und Springer 2004).

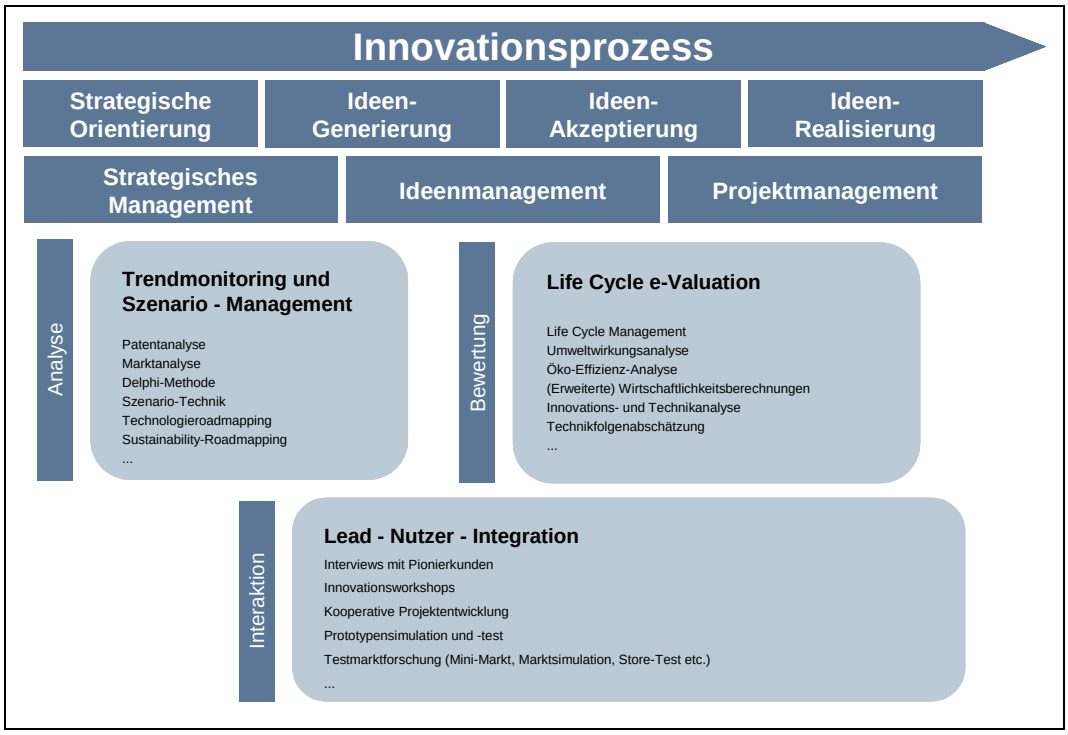

Abbildung 2-1: Der Innovationsprozess

2.1 Trendmonitoring im Szenario-Management

Immer kürzer werdende Innovationszyklen können nur beherrscht werden, wenn die für den Innovationsprozess benötigten Informationen jederzeit unmittelbar zur Verfügung stehen. Dazu ist sowohl die dauerhafte Verfolgung gesellschaftlicher und technologischer Entwicklungen, als auch die ständige Beobachtung von Märkten und Wettbewerbern notwendig. Die Identifikation, die dauerhafte Überwachung von Indikatoren und Deskriptoren und die Erstellung von Szenarien, die Veränderungen im Beobachtungsfeld signalisieren, ist die Kernaufgabe des Trendmonitorings. Dabei ist die isolierte Beobachtung einzelner Indikatoren und Deskriptoren selten sinnvoll; vielmehr müssen diese in ihrer Gesamtheit unter Berücksichtigung der gegenseitigen Wechselbeziehungen betrachtet werden. Hierzu können zum Beispiel Szenarien verwendet werden, die verschiedene Indikator-Ausprägungen in unterschiedlichen Kombinationen repräsentieren.

Für die Entwicklung von Szenarien kann die Szenario-Technik (vgl. Reibnitz 1991) eingesetzt werden. Sie ist eine Planungstechnik, mit der mehrere, deutlich unterschiedliche Szenarien erstellt werden können. Als Erweiterung der Szenario-

Technik kann auch das Szenario-Monitoring Anwendung finden. Dabei werden Szenarien nicht nur einmalig erstellt, sondern über einen längeren Zeitraum hinweg kontinuierlich angepasst.

2.2 Life Cycle e-Valuation

Life Cycle e-Valuation bezeichnet die Bewertung von Umweltwirkungen durch informationstechnische Unterstützung bzw. Online-Tools als einem zentralen Bestandteil einer Nachhaltigkeitsbewertung von Produkten, Dienstleistungen oder Produktsystemen. Sie reiht sich damit in umfassende Konzepte wie das Life Cycle Management (LCM) oder die Integrierte Produktpolitik ein.

Von zentraler Bedeutung sind dabei folgende Methoden:

- Umweltwirkungsbewertung/Ökobilanzierung

  Durch die Ökobilanzierung (engl. LCA – Life Cycle Assessment) können die Umweltwirkungen eines Produkts oder einer Dienstleistung bewertet werden. Dabei wird im Allgemeinen der gesamte Lebenszyklus des Produkts/der Dienstleistung betrachtet und die dabei fließenden Stoff- und Energieströme mit deren Umweltwirkungen bilanziert und beurteilt. Um die Methode handhabbarer für die Praxis zu machen, wurden vereinfachte Methoden der Ökobilanz entwickelt, die ebenfalls für die Umweltwirkungsbewertung genutzt wer-
  4den können.

- Erweiterte Wirtschaftlichkeitsrechnungen

  Unter dem Begriff *erweiterte Wirtschaftlichkeitsrechnung* werden konventionelle Methoden der Wirtschaftlichkeitsrechnungen für den betrieblichen Einsatz verstanden, die um ökologische Bewertungsmaßstäbe erweitert wurden. Dazu gehören unter anderem Ansätze wie die Ökoeffizienzanalyse, das Life Cycle Costing oder das Design for Environment.

2.3 Lead User Integration

Lead User Integration ist ein Konzept im Bereich des Innovationsmanagements, das die Einbindung von Anwendern mit hohem Innovationsbedarf in den Produktentwicklungsprozess beschreibt. Wichtige Phasen des Ansatzes sind:

- die Identifikation von Trends in relevanten Suchfeldern,
- die Identifikation von Pionierkunden,
- die gemeinsame Entwicklung eines Produktkonzeptes sowie
- der Markttest der generierten Produktlösungen.

Wichtig bei der Entwicklung von komplexen Produkten ist die Auswahl derjenigen Pionierkunden, deren Pioniereigenschaft für das komplette Produkt gilt.

3 Information Workflow

Die Verwendung der Begriffe *Daten*, *Informationen* und *Wissen* gestützt durch unterschiedliche Definitionen kann zu unnötigen Missverständnissen führen. Aus diesem Grund sollten diese Begriffe hier kurz erläutert werden.

Daten sind die Repräsentation von Fakten, die von Menschen und/oder von Maschinen gelesen und interpretiert werden können. *Informationen* sind Daten die in einem bestimmten Kontext dargestellt werden, also von Menschen oder von Maschinen interpretiert werden. *Wissen* ist vom Menschen verarbeitete (verinnerlichte) Information, die zur Problemlösung und zur Erstellung neuen Wissens verwendet werden kann.

Die Lösung komplexer Aufgaben erfordert meistens die Aufteilung des Hauptproblems in möglichst abgrenzbare Teilprobleme. Für die Definition von Teilproblemfolgen und für die Steuerung der Lösungsreihenfolge kann der so genante Verarbeitungsfluss verwendet werden.

Der Verarbeitungsfluss (*Workflow*) ist eine Folge von Verarbeitungsschritten. Die einzelnen Verarbeitungsschritte erhalten Daten und/oder Informationen als Eingabe und, geben diese, nach geeigneter Transformation, an Folgeprozesse weiter. Hierbei ist die Reihenfolge der Verarbeitungsschritte maßgebend.

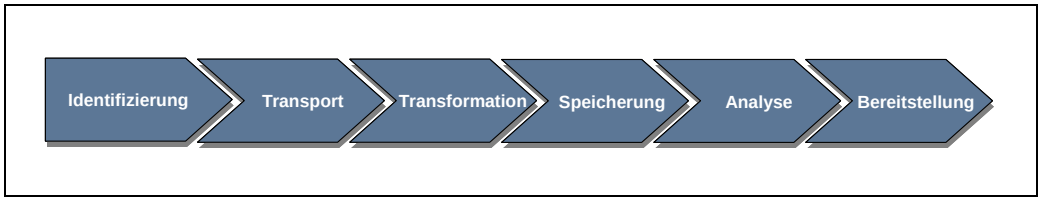

Abbildung 3-1: Information Workflow

Wie bereits angedeutet, stellt die Verarbeitung großer Daten- und Informationsmengen für Unternehmen unserer Zeit eine besondere Herausforderung dar. Vielfältige Informationen, wie Produktionsdaten oder Informationen über neue Technologien sind für die bewusste Gestaltung des Innovationsprozesses notwendig. Daten können sowohl aus externen Datenquellen, als auch aus den Unternehmen selbst stammen. Unabhängig davon müssen sie zur richtigen Zeit in geeigneter Form als Entscheidungsgrundlage bereitgestellt werden. Dafür ist eine zielgerichtete Informationsverarbeitungsstrategie notwendig, die den Verarbeitungsprozess in sinnvolle Phasen unterteilt. Die einzelnen Phasen bilden den Verarbeitungsfluss, den so genannten *Workflow*.

Da wir uns hauptsächlich mit der Verarbeitung von Informationen beschäftigen, verwenden wir in diesem Zusammenhang den Begriff *Information Workflow*. Die

einzelnen Phasen des Workflows sind in Abbildung 3-1 abgebildet und werden in Abschnitt 3.1 erläutert.

3.1 Phasen der Informationsverarbeitung

Wie im vorangehenden Abschnitt beschrieben, ist die komplexe Aufgabe der Informationsverarbeitung im Bereich des Innovationsmanagements in mehrere Phasen aufgeteilt. Die nächsten Abschnitte erläutern die einzelnen Informationsverarbeitungsphasen und verbinden sie mit spezifischen Innovationsaufgaben.

3.1.1 Identifizierung

Im Laufe eines Innovationsprojektes müssen vielfältige Entscheidungen auf Grundlage diverser Informationen getroffen werden. Die Identifizierung der richtigen Informationen und Informationsquellen stellt oft eine schwierige Aufgabe dar. Aus diesem Grund sind Techniken und Technologien für die Identifizierung von Informationen und Informationsquellen für das Innovationsmanagement essenziell. Im Themenfeld *Trendmonitoring im Szenario-Management* sind in erster Linie Fakten relevant, die die Einschätzung von ermittelten Einflussfaktoren für eine bestimmte Leitfrage erleichtern. Im Bereich *Life Cycle e-Valuation* stehen umweltpolitische Gesetze, Meldungen und Diskussionen aktueller Fragestellungen mit allgemeinem Bezug zu einem Produkt und/oder einer Branche im Mittelpunkt der Betrachtung, während im Themenfeld *Lead User Integration* Pionierkunden gesucht werden, die in den Entwicklungsprozess eingebunden werden können.

Die Identifizierung von Daten- und Informationsquellen gehört, neben der Analyse dieser Quellen, zu den schwierigeren Aufgaben im Information Workflow. Informationen sind meist in vielen verschiedenen heterogenen Informationsquellen zu finden. Heterogene Daten müssen vereinheitlicht, und in ein Gesamtschema integriert werden, um die Weiterverarbeitung zu ermöglichen.

Unternehmensintern sind Daten und Informationen oft in Dokumenten in einem Dateisystem, in Dokumentmanagementsystemen (vgl. Abschnitt 4.4.2) oder in Datenbanken (vgl. Abschnitt 4.4.1) zu finden. Während die Suche nach Inhalten in einer Datenbank oder in einem Dokumentmanagementsystem zu den Hauptmerkmalen zählt, und damit in aller Regel sehr gut umgesetzt ist, verspricht die Suche nach Informationen in großen Dokumentenbeständen im Dateisystem wenig Erfolg. Die Entwicklung und Implementierung von Suchfunktionen, die nicht nur die Syntax, sondern auch die Semantik der Daten berücksichtigen, ist Gegenstand aktueller Forschungsarbeiten. Nach unserem Wissensverständnis, steckt Anwendungs- und Expertenwissen immer in den Köpfen von Personen. Demnach können menschliche Wissensträger nur in Expertenverzeichnissen, oder durch weitere Personen, die Experten kennen, identifiziert werden.

Die erwähnten Informationsquellen können selbstverständlich auch unternehmensextern (z.B. bei anderen Unternehmen) existieren. Zusätzlich sollte das World Wide Web, das sich in den letzten Jahren zur größten aber auch zur heterogensten Informationsquelle der Menschheit entwickelt hat, in die Liste der externen Quellen aufgenommen werden. Insbesondere sind auch die so genannten „Deep Web“ Ressourcen zu beachten, die große Datenmengen, welche durch Portale zugänglich gemacht werden, darstellen.

Während die Suche und die Identifikation von Informationen in strukturierten Datenmengen schon oft diskutiert und mit zahlreichen Anwendungssystemen umgesetzt wurden, sind im Bereich des World Wide Webs diese Aufgaben Gegenstand aktueller Forschung. Die für das Innovationsmanagement relevanten Technologien aus diesem Bereich stellen wir in Abschnitt 4.1 vor.

3.1.2 Transport

Vor wenigen Jahren speicherten Unternehmen den Großteil ihrer Daten noch in gedruckter Form. Der Transport dieser Daten war oft umständlich und kostenintensiv. Im Zeitalter der computerunterstützten Datenverarbeitung liegt im Gegensatz dazu ein Großteil dieser Daten in elektronischer Form vor. Der Datenaustausch in unternehmensinternen Netzwerken oder im Internet ist günstiger und vor allem sehr viel schneller. In vielen Fällen zählen die Kommunikation und der Transport von Daten im Rahmen des Informationsverarbeitungsprozesses zu den einfach zu lösenden Aufgaben.

Ein besonderes Augenmerk muss jedoch auf den Datentransport in elektronischen Netzen gelegt werden, wenn sicherheitskritische Daten transportiert werden sollen. Im Bereich des Innovationsmanagements ist dies häufig der Fall. Informationen über neuen Technologien, Herstellungsprozesse oder Produktideen dürfen keinesfalls außerhalb definierter organisatorischer Einheiten sichtbar werden, und müssen deshalb sicher transportiert werden. Basistechnologien für den sicheren Transport von Daten stellen wir in Abschnitt 4.2 vor.

3.1.3 Transformation

Daten aus externen Informationsquellen liegen selten in einer Form vor, die optimal der unternehmensinternen Weiterverarbeitung angepasst ist. Dies gilt sowohl für die Struktur als auch für den Inhalt. Oft ist ein Teil der erworbenen Datenmenge fehlerhaft, im falschen Format oder schlicht uninteressant. Fehler müssen, wenn möglich, korrigiert werden; nicht beseitigbare Fehler und unnötige Daten müssen aus dem Datenbestand herausgefiltert werden. Ein ebenfalls oft notwendiger Schritt ist die Anpassung des Datenschemas an die unternehmensinternen Anforderungen. Erst danach ist die Integration der Daten in die bestehende Informationsinfrastruktur des Unternehmens möglich. Diese Aufgaben bezeichnen wir zusammenfassend mit dem Begriff *Datentransformation*.

Die Notwendigkeit des Datenaustausches sowie Komplexität und Kosten der Datentransformation führten dazu, dass in vielen Bereichen mittlerweile standardisierte Datenaustauschformate verfügbar sind. Der XML-Standard bildet häufig die Grundlage für solche Austauschformate, den eigentlichen Datenaustausch und die damit verbundenen Transformationen. In Abschnitt 4.3 werden relevante Technologien für die Transformation von Daten vorgestellt.

3.1.4 Speicherung

Nachdem Daten in eine, für die Weiterverarbeitung geeignete Form, transformiert wurden, müssen sie gespeichert werden. Die Speicherstrategie hängt natürlich von der gewählten Datenform und vom Datenschema ab. Sind zum Beispiel viele elektronische Dokumente zu speichern, bietet sich die Möglichkeit an, diese in einem Dokumentmanagementsystem (vgl. Abschnitt 4.4.2) oder in einer digitalen Bibliothek (vgl. Abschnitt 4.4.3) abzulegen. Für die Speicherung von strukturierten Daten sollten Datenbanksysteme eingesetzt werden.

Die gewählte Transformations- und Speicherstrategie bestimmt maßgeblich die Weiterverarbeitungsmöglichkeiten der gespeicherten Daten. Müssen z.B. während der Verarbeitungsphase hohe Datendurchsatzraten erreicht und komplexe Abfragen gestellt werden können, dann führt kein Weg an einem Datenbanksystem mit einer geeigneten Abfragesprache vorbei.

Wie bereits im vorangehenden Abschnitt erwähnt, liegen Daten immer öfter in, Austauschformaten vor, die auf XML basieren. Es liegt nahe, XML-Daten direkt in speziellen XML-Datenbanken (vgl. Abschnitt 4.4.1) zu speichern. Die Verarbeitung und die Präsentation werden dann durch die etablierten XML-Abfrage- und XML-Transformations-Sprachen unterstützt.

3.1.5 Analyse

Wie bereits erwähnt, müssen die Datenspeicherung und die Weiterverarbeitung aufeinander abgestimmt werden. Das effiziente Auffinden von Informationen muss durch eine geeignete Suchfunktion mit ausreichend mächtiger Abfragesprache ermöglicht werden.

Die Analyse von Daten umfasst nicht nur die inhaltsbezogene Suche sondern auch die Aggregation und das Filtern von Daten sowie die Identifikation von Mustern und Zusammenhängen. Für spezifische Problemstellungen müssen spezielle Anwendungen und Tools entwickelt werden. Dabei können verschiedene Entwicklungsumgebungen, Programmiersprachen sowie mathematische und statistische Verfahren zum Einsatz kommen. Die vollständige Auflistung dieser Methoden und Technologien ist in diesem Dokument jedoch nicht zweckmäßig.

Werden allerdings strukturierte Datenmengen in Datenbanken gespeichert und verspricht die Analyse der Daten neue Erkenntnisse, die bedingt durch die große Datenmenge nicht von Menschen erarbeitet werden können, dann bieten sich sogenannte Data Warehouses als integrierte Datengrundlage sowie OLAP- und Data-Mining-Technologien für die Analyse an. Wir stellen diese Möglichkeit der Datenverarbeitung in Abschnitt 4.5 vor.

3.1.6 Bereitstellung

Nach der Analysephase müssen die Daten anderen Systemen und Menschen zugänglich gemacht werden. Die Weitergabe der Daten für eine anschließende maschinelle Bearbeitung durch Fremdsysteme ist vergleichsweise unkompliziert, da höchstens einfache Transformationen notwendig sind. Dagegen stellt die Informationspräsentation für menschliche Betrachter immer eine Herausforderung dar. Dabei spielt eine Vielzahl von Faktoren, wie Softwareergonomie, Wissensstand des Benutzers, etc. eine Rolle.

In Kapitel 4.6 werden Technologien vorgestellt, die vor allem im Bereich des World Wide Webs für die Bereitstellung von Informationen eingesetzt werden können. Sie unterstützen die ansprechende Darstellung von Daten und die Erstellung von ergonomischen Benutzungsschnittstellen.

4 Allgemeine Technologieübersicht

Auf den folgenden Seiten soll eine Übersicht über die derzeit im Umfeld des Internet eingesetzten Methoden und Technologien zur Be- und Verarbeitung von Informationen gegeben werden. Zudem wird aufgezeigt, welche neuen Standards und Entwicklungen derzeit das Geschehen in der Informationsverarbeitung bestimmen und worauf sich die in sie gesetzten Hoffnungen stützen. Diese Technologieübersicht ist entlang der Phasen des Information Workflows gegliedert. Sie bildet die Grundlage für die Bearbeitung spezieller Fragestellungen, die im Zentrum der einzelnen Themenfeldern des Forschungsprojekts *nova-net* stehen.

4.1 Technologien für die Identifizierung

Mit der Identifizierung von Daten und Informationen beginnt die eigentliche Informationsverarbeitung. Dabei sind drei Fragestellungen von Bedeutung: woher kommen potenzielle Informationen, wie lassen sie sich möglichst automatisch identifizieren und wie in den Workflow integrieren.

4.1.1 Informationsquellen

Das Internet präsentiert sich dem Nutzer heute als ein Geflecht aus unzähligen öffentlich und nicht öffentlich zugänglichen heterogenen Datenquellen, die mittels unterschiedlicher Dienste und Protokolle nutzbar sind. Der derzeit wichtigste dieser Dienste ist dabei sicherlich das *Word Wide Web (WWW)*, da es als Schnittstelle zu verschiedenen Arten von Informationssystemen fungiert, bzw. selbst als eine Art Datenspeicher angesehen werden kann. So werden neben einfachen Informations- und Linksammlungen auch komplexe und interaktive Datenbanken wie Patentinformationssysteme oder Bibliotheksrecherchesysteme über eine einzige Komponente des Internets, das WWW, integriert. Da mittlerweile die kostengünstig verfügbare Speicherkapazität in einem solchen Umfang zugenommen hat, dass Inhalte kaum noch gelöscht werden müssen, sondern archiviert und für lange Zeit verfügbar gehalten werden können, bietet sich den Nutzern eine permanent wachsende Daten- und Informationsbasis, auf die sie jederzeit nach Belieben zugreifen können. Die Nutzung des WWW als Informationsbasis setzt zunächst die Identifizierung der im jeweiligen Kontext benötigten Informationen voraus. Dieses wird in den folgenden Abschnitten thematisiert. Darüber hinaus kann die Aktualität der verfügbaren Information häufig nicht eindeutig bestimmt werden und es existieren keinerlei Garantien, dass einmal identifizierte Dokumente und Informationen im WWW für einen bestimmten Zeitraum zur Verfügung stehen. Nutzer und Anwendungen, die das World Wide Web als Daten- und Informationsbasis einsetzten, müssen die genannten Einschränkungen berücksichtigen.

Aufgrund der Fülle der Informationen und der verschiedenen Ansätze seitens der Content-Anbieter, diese zu strukturieren, ist die Aufgabe der Identifizierung rele-

vanter Informationen im WWW mitnichten trivial. Erschwerend kommt hinzu, dass vor allem bei HTML-Seiten, die den Großteil aller Web-Dokumente ausmachen, die multimediale Darstellung der Inhalte zunehmend in den Vordergrund rückt. Das Resultat ist, dass ihre HTML-Beschreibungen zu großen Teilen aus Formatierungselementen bestehen, die bestenfalls keinen inhaltlichen Wert besitzen, zudem jedoch oft eine automatische Verarbeitung der Inhalte deutlich erschweren, indem die Struktur des Inhalts zugunsten des Layouts teilweise aufgegeben wird. Dies gilt insbesondere für den verstärkten Einsatz von Tabellen und Flash-Animationen.

Die Identifizierung von potenziell relevanten Informationen in einem Datenpool ist deshalb eine der komplexeren Aufgaben im Rahmen des Information Workflow. Das Hauptproblem liegt in der Schwierigkeit, natürliche Sprache, in welcher sich der überwiegende Teil aller Webseiten präsentiert, maschinell zu verarbeiten. D.h. es können zwar Begriffe aus dem geschriebenen Inhalt extrahiert werden, jedoch existiert keine allgemein anwendbare Methode aus den Begriffen und ihrem textuellen Kontext eine Bedeutung zu ermitteln. Suchmaschinen versuchen der Informationsflut zu begegnen, indem sie das Web nach öffentlich zugänglichen Datenquellen durchsuchen und diese indizieren, um dem Benutzer anschließend ein Auffinden der Daten durch Eingabe von Suchbegriffen und Themengruppen zu ermöglichen. Anhand verschiedener Kriterien werden die Ergebnisse von Suchanfragen klassifiziert und nach ihrer angenommenen Relevanz bewertet, um dem Benutzer eine Vorauswahl zu präsentieren und ihn von der Sichtung potenziell irrelevanten Materials zu entlasten. Ein Problem dieser Vorgehensweise besteht darin, dass kaum Anfragen zu realisieren sind, die über die syntaktische Analyse der Datenbasis hinausgehen. Darüber hinaus sind die verwendeten Relevanz-Maße für den Nutzer in der Regel intransparent und können einer kommerziell motivierten Beeinflussung unterliegen.

4.1.2 Metadaten zur Beschreibung von Inhalten

Aufgrund dieser Schwierigkeiten wird seit einiger Zeit der Begriff bzw. die Methodik des *Semantic Web* propagiert. Im Rahmen des Semantic Web existieren verschiedene Ansätze, um, zusätzlich zu den reinen Daten, Informationen über deren Inhalt in maschinell verarbeitbarer Form zu speichern. Dies geschieht meist anhand von separaten oder integrierten Metadaten, also Daten, welche die eigentlichen Informationen beschreiben. Ein zentraler Baustein solcher Meta-Daten im Rahmen des Semantic Web ist das *Resource Description Framework (RDF)*, ein Standard, der es erlaubt, beliebige Internet-Ressourcen mit prinzipiell frei definierbaren Eigenschaften zu charakterisieren (vgl. W3C 1999 sowie Fensel 2002). Prinzipiell frei definierbar bedeutet, dass zwar jedermann beliebige Eigenschaften festlegen kann, diese aber erst dann einen höheren Nutzen ergeben, wenn sie eine gewisse allgemeine Akzeptanz erfahren haben und auch verbreitet eingesetzt werden. RDF-Datensätze setzen sich dabei aus drei Komponenten zusammen – Subjekt, Prädikat und Objekt, wobei das Prädikat die Verbindung zwischen Subjekt und Ob-

jekt beschreibt und damit den eigentlichen informationellen Mehrwert liefert. Da RDF selbst keine Prädikate definiert, sondern lediglich ein Framework für die Definition von Metadaten bereitstellt, sind konkrete Ausprägungen in Form sog. RDF-Vokabulare gefordert. Ein Beispiel hierfür ist der *Dublin-Core-Standard* (vgl. RFC-2413 1998), einer Sammlung von Meta-Definitionen anhand derer Angaben über Rahmendaten wie Autor, Datum der Veröffentlichung, Lizenzbedingungen, etc. einer Internet-Ressource gemacht werden können (vgl. Abbildung 4-1). RDF-Vokabulare lassen sich mittels RDF-Schema spezifizieren. Dabei kommen objektorientierte Konstrukte wie z.B. Aggregationen, Vererbungsbeziehungen oder Wertebereichseinschränkungen zum Einsatz, womit sich komplexe Szenarien beschreiben lassen.

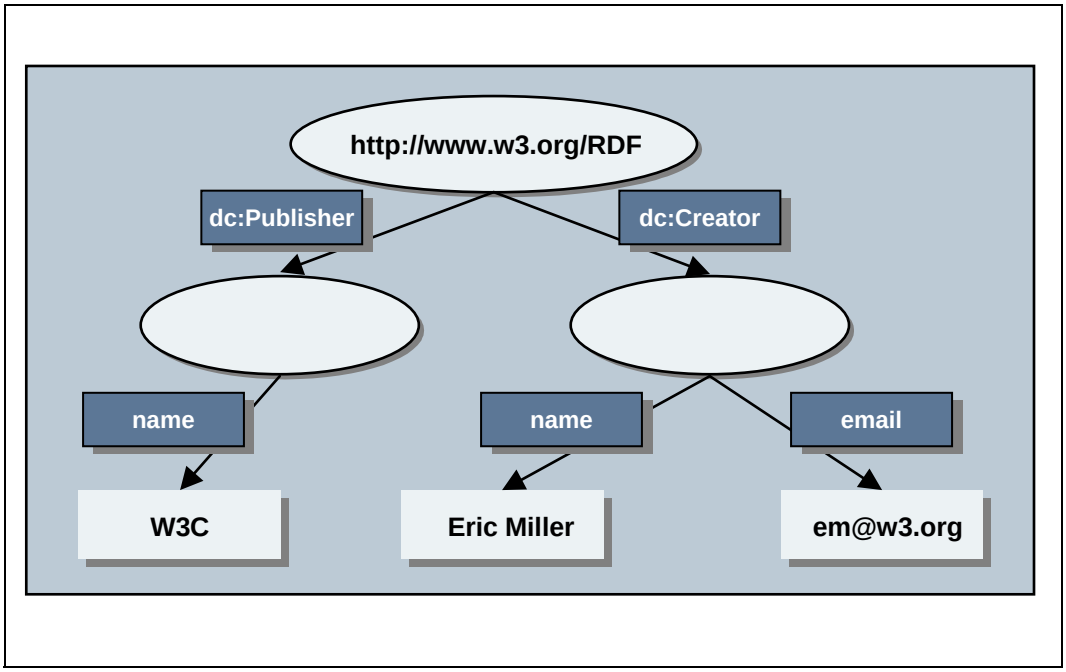

Abbildung 4-1: Beispielhafter RDF-Graph zur Beschreibung einer Web-Site

4.1.3 Wissensverarbeitung mittels Ontologien

Einen weitergehenden Ansatz stellen Ontologien dar. Mittels Ontologien können Objekte einer realen oder virtuellen Welt beschrieben werden sowie Zusammenhänge zwischen diesen Objekten oder ihren Beschreibungsklassen erfasst und zueinander in Relation gesetzt werden. Der Mehrwert von Ontologien gegenüber rein metasprachlichen Beschreibungen liegt darin, dass mittels sog. Inferenzmaschinen die durch die Ontologie definierten Beziehungen verarbeitet werden können, um neues, nicht explizit vorhandenes Wissen daraus abzuleiten. Die modellierbaren Beziehungen folgen dabei logischen Konstrukten wie Äquivalenz- und Transitivitätsbedingungen und erlauben somit eine automatisierte Verarbeitung

von Anfragen gegen eine Wissensbasis. Damit lässt sich auch unvollständiges Wissen über Zusammenhänge ausdrücken und nutzen.
Das Standardisierungsgremium des Internets (W3C) hat mit *OWL* (vgl. W3C 2004a) einen Standard auf den Weg gebracht, der Verfahren spezifiziert, die es ermöglichen, solche Ontologien zu definieren und mit ihnen zu arbeiten. OWL basiert dabei auf RDF und erweitert dieses um eine formale Semantik. Bezüglich der Mächtigkeit besitzt OWL verschiedene Untermengen - *OWL Lite*, *OWL DL* und *OWL Full*, die unterschiedlich komplexe logische Verarbeitungsschritte ermöglichen. Dabei besitzt OWL Full die größte Mächtigkeit, jedoch kann nicht für jede Anfrage eine Garantie bezügliche Berechenbarkeit und Entscheidbarkeit gegeben werden. Oftmals genügen aber auch die weniger komplexen Untermengen OWL DL bzw. OWL Lite, bei denen einige Konstrukte nicht enthalten sind oder nur unter bestimmten Bedingungen eingesetzt werden können. Beispielsweise lassen sich Klassifikationsprobleme in der Regel bereits mittels OWL Lite lösen.

Insbesondere der Ontologie-Ansatz scheint viel versprechend und zukunftsträchtig zu sein, da er als der Schlüssel zur automatischen Verarbeitung von Inhalten angesehen werden kann (vgl. Fensel 2002). Problematisch ist und bleibt jedoch, dass die Möglichkeit zur Automatisierung nur dann gegeben ist, wenn der jeweilige Content-Anbieter einen entsprechend hohen Aufwand in die Auszeichnung von Inhalten investiert. Es ist nicht zu erwarten, dass dies für einen nennenswerten Anteil aller Web-Ressourcen geschehen wird, sondern nur in eng umgrenzten Bereichen, in denen sich die Anbieter entsprechende Vorteile von dieser Zusatzarbeit erhoffen können. In diesen Bereichen jedoch ergeben sich daraus große Potenziale.

4.1.4 Information Retrieval

Die oben genannten Methoden und Technologien haben zum Ziel, das Auffinden und die Identifizierung von Informationen in verteilten Datenbasen zu erleichtern. Dabei stehen, wie bereits angeführt, mit Suchmaschinen oder elektronischen Katalogen verschiedene Technologien zur Verfügung, um dies zu unterstützen. Diese werden unter dem Begriff des *Information Retrieval (IR)* subsumiert (vgl. Ferber 2003).

Die intuitivsten und am weitesten verbreiteten Verfahren basieren dabei auf der syntaktischen Analyse von Texten. Hierbei werden Dokumente von einem System analysiert und mit den darin gefundenen Begriffen indexiert. Die so aufgebauten Indexstrukturen ermöglichen eine schnelle Suche nach Dokumenten, die vom Benutzer angeforderte Stichworte enthalten. Zusätzlich können boolesche Ausdrücke, wie z.B. „(A oder B) und nicht C“, relativ einfach durch Mengenoperationen auf den von den Indexstrukturen gelieferten Ergebnismengen realisiert werden. Existiert zusätzlich zu den syntaktischen Informationen auch Wissen über den Inhalt der Dokumente, etwa durch metasprachliche Auszeichnungen, so können zudem Klassifikationsverfahren eingesetzt werden. In diesem Fall erzeug das IR-

System aus der zu analysierenden Dokumentenmenge verschiedene Dokumentenklassen, denen jedes einzelne Dokument zugeordnet werden kann. Diese Klassen wiederum können bei Bedarf selbst strukturiert (z.B. hierarchisch) angeordnet werden, wodurch sich ein Klassifikations- und Beziehungsnetz unter den Dokumenten ausbildet. Beispiele solcher Klassifikationen existieren im WWW mit *Yahoo* (www.yahoo.de) oder dem *Open Directory Project* (www.dmoz.org).

Anfragen eines Benutzers sind oft unscharf oder ungenau, d.h. vorhandenen Informationen können mit den oben genannten Methoden nicht gefunden werden, weil die Fragestellung des Benutzers ein System voraussetzt, das Wissen und ein zumindest oberflächliches Verständnis der verwalteten Daten besitzt. Zum Teil ist dieses Verständnis mit genannten Klassifikationen bereits erreicht, bessere Ergebnisse lassen sich jedoch erzielen, wenn zusätzliche Methoden zum Einsatz kommen, die auch Dokumente berücksichtigen, welche vom Benutzer *wahrscheinlich* als relevant betrachtet werden. Hier können beispielsweise semantische Netze (vgl. Abschnitt 4.1.3 über Ontologien) und Thesauren eingesetzt werden. Ein Thesaurus ist eine Sammlung von Bezeichnungen und eindeutig zugeordneten Begriffen. Die in einem Thesaurus festgehaltenen Relationen zwischen Begriffen erlauben es, begriffliche Verwandtschaften zu ermitteln.

Dennoch setzen auch derartige Systeme voraus, dass der Benutzer weiß, wonach er sucht. Oftmals ist dies jedoch nicht der Fall bzw. die vorhandene Datenbasis eignet sich nicht für derartige Suchanfragen. Mit Data-Mining-Verfahren (vgl. Abschnitt 4.5.2) lässt sich solchen Problemen begegnen, indem strukturierte Datenbestände nach Regelmäßigkeiten und Mustern durchsucht werden. Die Ergebnisse dieses *Minings* werden dem Benutzer präsentiert, und können als Basis für weitere Suchen oder Anfragen an eines der oben genannten Systeme dienen.

4.2 Technologien für den Transport

Die Grundzüge des Datentransports in Computernetzen werden häufig anhand des ISO/OSI-Netzwerkmodells beschrieben (vgl. ISO 1984 sowie Rose 1990). Auch wenn sich die Praxis nicht streng nach diesem Modell richtet, so lassen sich doch die unterschiedlichen Aspekte des Datentransports anhand dieses Modells erläutern. Demnach sind sieben aufeinander aufbauende Ebenen zu unterscheiden (Abbildung 4-2, linke Seite).

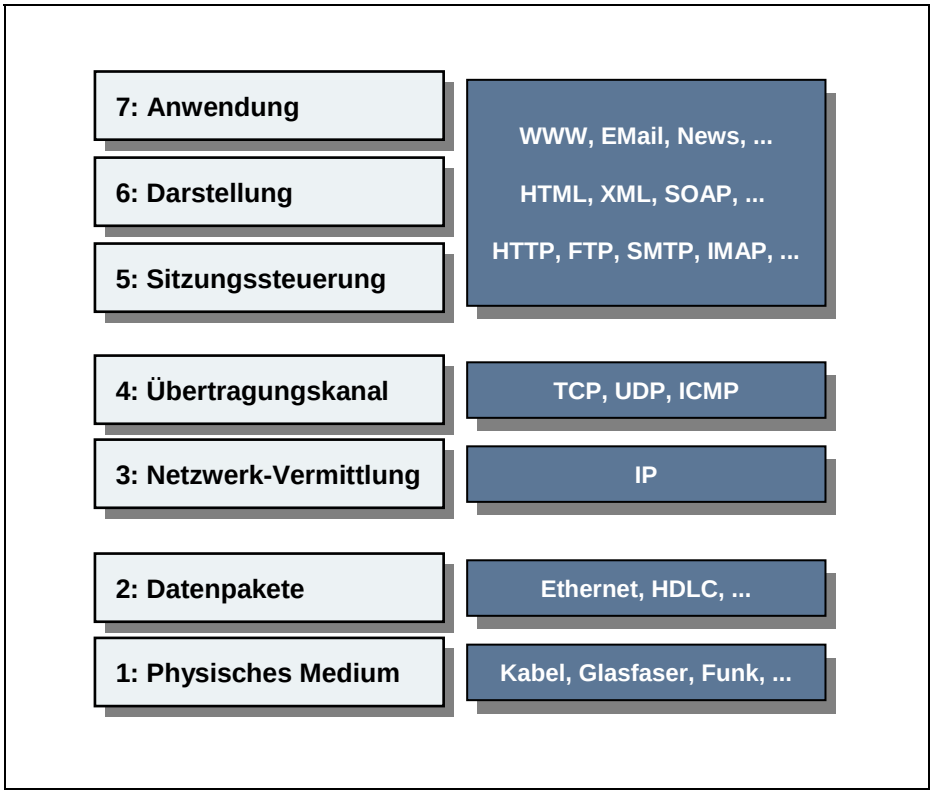

Abbildung 4-2: OSI-Netzwerkmodell und aktuelle Protokolle/Anwendungen

4.2.1 Basisprotokolle zur Datenübertragung

Die unterste Ebene (1) stellt die physische Verbindung dar, die zur Kommunikation zweier Partner genutzt wird. Klassische Medien hierfür sind TwistedPair-Kabel, Glasfaser und zunehmend auch Funkverbindungen.
Auf der zweiten Ebene erfolgt die Aufteilung der zu übertragenden Daten in Pakete. Hier werden bereits einfache Methoden zur Verifikation der Korrektheit der empfangenen Daten bzw. zur Korrektur von Übertragungsfehlern eingesetzt. Sind mehrere potenzielle Kommunikationspartner an das verwendete Übertragungsmedium angeschlossen, so erfolgt in Ebene zwei auch die Adressierung des jeweils aktuellen Partners. Beispiele für hier eingesetzte Protokolle sind Ethernet, Token Ring oder HDLC.
Findet auf den unteren beiden Ebenen ausschließlich Kommunikation zwischen Partnern eines physischen Netzes statt, so wird in der dritten Ebene eine Koppelung von unabhängigen Netzwerken realisiert. Somit können zwei Partner auch dann miteinander kommunizieren, wenn keine direkte physische Verbindung zwischen ihnen besteht, aber Brücken zwischen ihren Netzwerken existieren (sog. Router).
In der darauf aufbauenden vierten Ebene, der Transportebene, wird ein Datenkanal zwischen zwei Kommunikations-Endpunkten aufgebaut. Die hier eingesetzten

Protokolle stellen die Integrität der übertragenen Datenpakete sicher (in der Regel wieder durch Prüfsummenberechnung) und setzen diese in der richtigen Reihenfolge wieder zusammen. Letzteres wird klar vor dem Hintergrund, dass zwei aufeinanderfolgende Datenpakete auf der dritten Ebene durchaus andere Wege nehmen können. Dies ist z.B. im Falle von teilweisen Netzüberlastungen oder –ausfällen notwendig. Zudem wird je nach Protokoll die Vollständigkeit der übermittelten Daten garantiert, d.h. es kann z.B. automatisch eine erneute Übertragung eines verlorenen gegangenen Datenpaketes veranlasst werden. Die Ebenen drei und vier werden im Internet in aller Regel durch die beiden Protokolle *Transmission Control Protocol* und *Internet Protocol (TCP* bzw. *IP)* implementiert.
Die ersten vier Ebenen werden als „End-To-End"-Services bezeichnet und decken den Transport von Daten zwischen zwei oder mehreren Kommunikationspartnern ab. Syntax und Semantik der übertragenen Daten, also die eigentlichen Informationen, sind auf diesen Ebenen allerdings ohne Bedeutung und werden erst in den höheren Ebenen verarbeitet.

4.2.2 Höhere Protokolle und Formate auf Anwendungsebene

Anwendungs- und damit informationsspezifische Protokollfestlegungen werden in den drei oberen Ebenen (fünf bis sieben) getroffen. Hier existiert in der Praxis meist keine scharfe Trennung mehr, es findet vielmehr eine Integration der Funktionalitäten zu einer Lösung mit ein bis zwei Protokollen statt. In diesem Rahmen geschieht die eigentliche Arbeit mit den übertragenen Daten, d.h. durch entsprechendes Kontextwissen extrahieren Anwendungen die durch die Daten repräsentierten Informationen und nutzen diese zur Weiterverarbeitung. Beispiele hierfür sind einerseits höhere Übertragungsprotokolle wie HTTP oder FTP, ebenso aber auch allgemeine Datenaustauschformate wie EDIFACT, XML, HTML, PDF oder Multimediaformate wie MPEG etc. Sie regeln zum einen den Ablauf des Datenaustausches (Dialogkontrolle), zum andern auch das Format sowie eine evtl. Darstellung der Informationen auf Ausgabegeräten. Die Sicherung der Daten durch Verschlüsselung und digitale Signaturen erfolgt ebenfalls in diesem Rahmen, indem zum einen der eigentliche Datenkanal verschlüsselt wird, und zum anderen durch den Einsatz kryptografischer Protokolle die Identität und Berechtigungen der Kommunikationspartner sichergestellt werden.

4.2.3 Datenaustauschformate

Zentral für den Austausch jedweder Informationen ist die Definition von gemeinsamen Schnittstellen zwischen den beteiligten Partnern. Dies ist die Aufgabe von Datei- bzw. allgemein Datenaustauschformaten. Zwar wird in diesem Zusammenhang XML immer populärer, dennoch ist in vielen Bereichen ein auf das jeweilige Anwendungsszenario zugeschnittenes Format vorzuziehen. Zum einen ist dies historisch bedingt, so haben sich z.B. bei den Bildformaten Standards wie TIFF und GIF schon seit Jahren etabliert und dementsprechend große Datenbestände wurden aufgebaut. Zum anderen gebieten oftmals Geschwindigkeits- und Spei-

cherplatzanforderungen den Einsatz von Formaten, bei denen im Gegensatz zu XML die Strukturbeschreibung der Daten nicht zwangsläufig mit diesen zusammen ausgeliefert wird. Sie kann dann implizit durch die Wahl des gewählten Formats gegeben sein, was bei der Übertragung in einer geringeren und schneller zu verarbeitenden Datenmenge resultiert. Aus dieser Motivation haben sich über die Jahre hinweg verschiedene Formate für unterschiedliche Anwendungen entwickelt und sind standardisiert worden bzw. haben den Status eines Quasi-Standards erreicht. Die Offenlegung dieser Schnittstellen ermöglicht aber erst den freien Austausch von Informationen über Anwendungsgrenzen hinweg. Eines der populärsten Beispiele hierfür ist das *Portable Document Format (PDF)* von Adobe, mittels welchem elektronische Dokumente auf nahezu jeder Betriebssystem- und Hardwareplattform betrachtet und verarbeitet werden können. Ähnliche Standards existieren auch in anderen Anwendungsbereichen wie z.B. bei Bildern und Grafiken (PNG, SVG, ...), der Videospeicherung (MPEG, AVI, ...) oder dem DXF-Format für technische Zeichnungen.

4.3 Technologien für die Transformation

Die Transformation von Daten kann im Information Workflow an verschiedenen Stellen auftreten (vgl. Abbildung 4-3). Zum einen handelt es sich um die Filterung von „neu erfassten“ Informationen mit dem Ziel, relevante Daten aus einer Informationsmenge zu isolieren und das Ergebnis dieses Prozesses zur weiteren Verarbeitung aufzubereiten und zu klassifizieren. Zum anderen erfolgt in der Regel erneut eine Filterung, wenn Benutzer über eine Präsentationsschicht auf gespeicherte Daten zugreifen. Letzteres ermöglicht die Verwendung einer Datenbasis unter verschiedenen Gesichtspunkten, indem Informationen kontextabhängig zusammengestellt bzw. gefiltert werden.
Offensichtlich findet auch in der Analysephase eine Informationsfilterung statt, ihr ist jedoch aufgrund der Komplexität ein eigener Abschnitt (4.5) gewidmet.

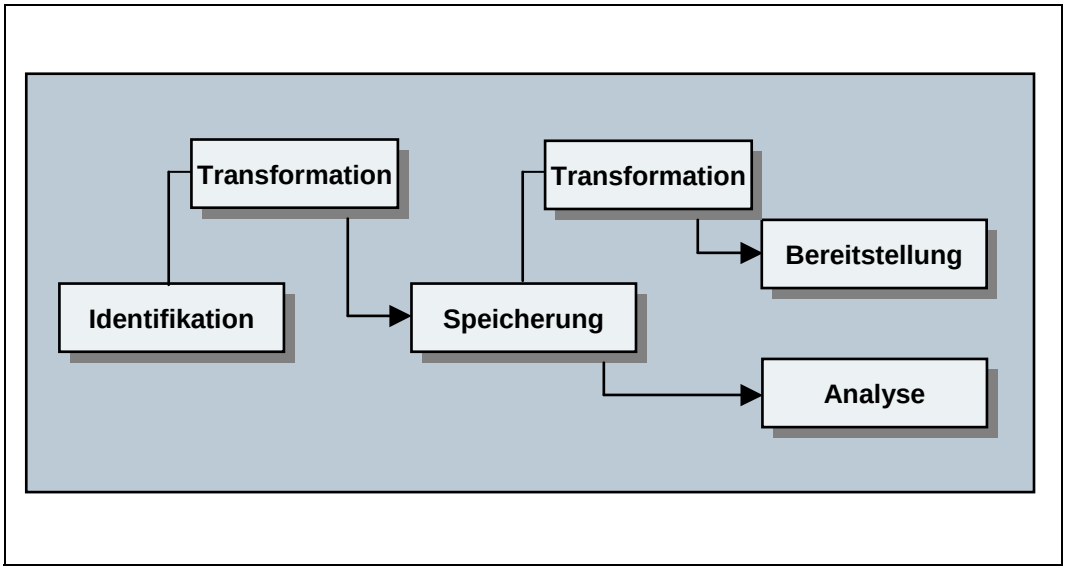

Abbildung 4-3: Filterung im Information Workflow

4.3.1 XML als Basisformat

Mit der *Extensible Markup Language (XML)* hat sich in den letzten Jahren eine Sprache zur Beschreibung von Datenformaten entwickelt und durchgesetzt, auf deren Basis der Austausch von Informationen zwischen beliebigen Partnern erleichtert wird. Zusätzlich entstanden verschiedenartige Technologien um diese Sprache herum, welche neben dem Austausch der hierarchisch strukturierten Daten auch deren Modifikation erlauben. Aufgrund der Mächtigkeit und Universalität von XML-basierten Technologien haben diese eine weite Verbreitung gefunden und können heute als Standard für den Informationsaustausch angesehen werden (vgl. Ferber 2003). In der Konsequenz wurden in den letzten Jahren für eine Vielzahl von Anwendungsbereichen Übereinkünfte über Datenformate auf der Basis von XML erzielt. Diese stellen für den jeweiligen Anwendungsbereich mittlerweile einen Standard oder Quasi-Standard dar und ersetzen spezialisierte, nicht XML basierte Datenformate bzw. werden zum Zweck des reibungslosen Datenaustausches eingeführt. Ein Beispiel hierfür ist *openTRANS*, das für den automatisierten Austausch von Geschäftsinformationen über das Internet und elektronische Marktplätze eingesetzt wird.

XML ist ein hierarchisches Datenformat, dessen Aufbau mit der Struktur eines Baumes vergleichbar ist. Ausgehend von einer Dokumentenwurzel verzweigt sich ein XML-Dokument in Äste, Unteräste und Blätter. Jeder dieser Knotenpunkte kann dabei neben seinen untergeordneten Teilbäumen zusätzliche Attribute beinhalten, die den Knoten näher beschreiben. XML wurde entwickelt um den reibungslosen Austausch von Daten zwischen Computerprogrammen zu ermöglichen. Daher ist auch die Struktur so gewählt, dass eine maschinelle Verarbeitung

und Prüfung möglichst einfach zu erreichen ist. Dennoch lassen sich XML-Dokumente auch von menschlichen Betrachtern in gewissem Maße verstehen und verarbeiten, da letztlich jegliche Information in Textform dargestellt wird.

Die am weitesten verbreiteten XML-Programmierschnittstellen sind in der *Simple API for XML (SAX)* und dem *Document Object Model (DOM)* spezifiziert. Sie erlauben auf einer relativ niedrigen Abstraktionsebene, Inhalt und Struktur von XML-Dokumenten zu analysieren bzw. zu modifizieren. Mit ihrer Hilfe können irrelevante Bereiche eines Dokuments auf einfache Art und Weise gelöscht bzw. übergangen oder spezielle Aktionen beim Lesen bestimmter Knoten angestoßen werden. Primär werden diese Technologien eingesetzt, wenn es um die direkte Manipulation eines XML-Dokuments geht, z.B. beim Entfernen eines Knotens inklusive aller Unterknoten oder beim Ändern eines Attributwertes. Sie bieten jedoch nur eine eingeschränkte Sicht auf den aktuell bearbeiteten Knoten und dessen direkte Nachbarn (über- bzw. untergeordnete Knoten sowie weitere Knoten derselben Hierarchieebene) und unterstützen eine globale Betrachtung nur unzureichend.

4.3.2 XML-Datenmanipulation mittels XSLT

Deutlich mächtigere Filtermöglichkeiten ergeben sich beim Einsatz von auf DOM und SAX aufbauenden Anfragesprachen und Transformationstechniken wie *XPath*, *XQuery* und *Extensible Stylesheet Language Transformations (XSLT)*. Mittels XPath lassen sich Teile eines XML-Dokuments durch eine Pfad-Angabe beschreiben, wobei der Wurzelknoten des Dokuments den Startpunkt des Pfades darstellt. Schritt um Schritt können dann Unterknoten oder Mengen von Unterknoten in die Spezifikation des Pfads aufgenommen werden, indem Kriterien für den Einschluss bzw. Ausschluss angegeben werden.
XSLT dient nun dazu, XML-Dokumente in unterschiedliche Zielformate und/oder Strukturen zu transformieren. Die Spezifikation der Transformation erfolgt dabei in einem sog. XSL-Stylesheet, in welchem anhand von XPath Knoten des Quelldokuments selektiert werden, zusammen mit den auf diesen Knoten auszuführenden Operationen (Ersetzungs- Lösch- und Einfügeoperationen, Iterationen, bedingte Verzweigungen, Sortierung, Rekursive Transformationsaufrufe etc.)
Auf diese Weise lassen sich sowohl nicht benötigte Informationen herausfiltern, als auch Struktur und Knoten- bzw. Attributnamen auf das Zielschema anpassen. Somit können gleichartige Daten aus unterschiedlichen Quellen in ein intern zur Verarbeitung verwendetes Schema transformiert werden, was nachfolgenden Verarbeitungsstufen den Datenzugriff erleichtert, da strukturelle Unterschiede der Quellen irrelevant werden und sich die gesamte Datenbasis über eine einzige Schnittstelle ansprechen lässt.

4.3.3 XML-Anfragesprache XQuery

Um Datenabfragen auf einem Dokument durchzuführen eignet sich prinzipiell XPath bzw. XPath in Verbindung mit XSLT. Dieses Vorgehen besitzt allerdings

zwei Nachteile: zum einen ist die XSL-Darstellung eher auf eine einfache maschinelle Verarbeitung denn auf gute Lesbarkeit durch Menschen ausgelegt. D.h. für den Entwickler/Benutzer ist der Umgang mit XSL nicht immer intuitiv, was wiederum einen gewissen Aufwand für Entwicklung und Wartung erfordert. Zum anderen enthält eine mit diesem Verfahren spezifizierte Anfrage bzw. Transformation relativ wenige Informationen über den Zusammenhang der einzelnen Teilschritte. D.h. es können zwar Ergebnisse erzeugt werden, die ausführende Komponente hat es aber schwer, Optimierungen am Ablauf vorzunehmen, da nur wenige zusätzliche Strukturinformationen über die Anfragesyntax transportiert werden können.

Anders verhält es sich dagegen mit der sich momentan noch in der Standardisierungsphase befindenden XML-Anfragesprache XQuery (vgl. W3C 2004b). Mit XQuery erhält der Anwender eine komfortable Anfragesprache für XML-Dokumente, die gegenüber der Verabeitung mit XSLT verschiedene Vorteile bietet. Zum einen bietet XQuery eine deutliche höhere Typsicherheit als XSLT/XPath. Damit kann deutlich mehr Semantik über die Sprache transportiert werden, was wiederum die Fehleranfälligkeit reduziert und die Verarbeitung beschleunigt. Ferner ist die Sprache an sich deutlich intuitiver und lehnt sich an den Prinzipien von gängigen Programmiersprachen an. Das wichtigste Prinzip dabei ist das FLWOR-Konzept (For-Let-Where-Order-Return), das Schleifeniterationen über Knotenmengen, Variablendefinition und –bindung sowie Filterung und Sortierung von Ergebnistupeln erlaubt. Letztgenannte Eigenschaften besitzt zwar XSLT ebenfalls, allerdings mit Einschränkungen hinsichtlich Einfachheit. XQuery hingegen baut auf einem komplexen Formalismus auf, der es erlaubt, dass weitreichende Transformationen und Optimierungen von Anfragen durch die Ausführungskomponente automatisch vorgenommen werden können. Damit ist XQuery bezüglich der Abfragemächtigkeit mit SQL, der Standard-Anfragesprache für relationale Datenbanken, vergleichbar (vgl. Abschnitt 4.4.1)

Zusammenfassend lässt sich sagen, dass XQuery sicherlich die Abfragetechnologie der Zukunft für XML-Daten sein wird. Neben seiner intuitiven Benutzbarkeit bietet es gleichzeitig hohes Potenzial für die Optimierung von Anfragen.

4.4 Technologien für die Speicherung

Alle informationsverarbeitenden Prozesse setzen eine stabile Speicherung der zugrunde liegenden Daten voraus. Datenbanksysteme stellen hierfür eine wichtige Basistechnologie dar. Der folgende Abschnitt greift darum verschiedene Aspekte der Datenbanktechnologie auf. Ergänzt wird diese Darstellung um jeweils einen Abschnitt zu Dokument-Management-Systemen und digitalen Bibliotheken. Gegenüber allgemeinen Datenbanksystemen bieten solche Systeme zusätzliche Funktionalität, die speziell auf die zu speichernden Inhalte zugeschnitten ist.

4.4.1 Datenbanktechnologie

Die persistente Speicherung von Daten in Datenbanksystemen ist für viele Anwendungsbereiche unabdingbar. Der Begriff *Datenbanksystem* umfasst hierbei sowohl den Datenbestand an sich als auch die Software zu dessen Verwaltung, das so genannte *Datenbankmanagementsystem (DBMS)*. Neben der persistenten Speicherung der Daten kontrolliert das DBMS unter anderem den gleichzeitigen Zugriff mehrere Nutzer auf einen zentralen Datenbestand, das Einhalten anwendungsspezifisch definierter Integritätsbedingungen sowie das Einhalten der administrationsseitig vorgegebenen Zugriffsrechte. Jedem Nutzer bzw. jeder Anwendung kann eine spezifische Sicht auf den Datenbestand zur Verfügung gestellt werden. Diese Dienste bietet das Datenbanksystem unabhängig von der einzelnen Anwendung an.

Ein wichtiger Vorteil der Nutzung von Datenbankmanagementsystemen ist die so genannte Datenabstraktion, d.h. eine möglichst weitgehende Unabhängigkeit der einzelnen Anwendungen von der Struktur, in der die Daten gespeichert werden. Um diese Abstraktion zu erreichen ist die Festlegung auf ein *Datenmodell* notwendig, das Speicher- und Implementierungsdetails vor den Datenbanknutzern verbirgt. Das gewählte Datenmodell bildet somit ein zentrales Unterscheidungsmerkmal für Datenbanksysteme (vgl. Elmasri und Navathe 2002).

Relationale Datenbankmanagementsysteme (RDBMS) basieren auf dem relationalen Datenmodell. Für den Anwender sind hierbei alle Daten als Zeilen in einer oder mehreren Tabellen sichtbar. Zur Manipulation der Daten in diesen Tabellen sowie zur inhaltsbezogenen Suche von Daten kommt die Anfragesprache SQL zum Einsatz (vgl. Melton und Simon 2002). Die den relationalen Datenbanksystemen zugrunde liegende Technologie hat sich in vielfältigen Anwendungsfeldern als geeignet erwiesen (Online Transaktionsverarbeitung, Data-Warehouse- und OLAP-Anwendungen, geografische Informationssysteme, …) und kann als ausgereift bezeichnet werden. Dies gilt insbesondere auch für die Verwaltung sehr umfangreicher Datenbestände.

Nichtsdestotrotz existieren Anwendungen, die den Einsatz anderer Datenmodelle nahe legen. So erlauben objektorientierte Programmiersprachen z.B. die Definition komplexer Datenstrukturen und die auf ihnen operierenden Funktionen (Methoden). Zur persistenten Speicherung und zum Lesen der im Programm verarbeiteten Daten ist eine explizite Konvertierung notwendig, die vom Datenmodell des zugrunde liegenden Datenbanksystems abhängt. *Objektorientierte Datenbankmanagementsysteme (OODBMS)* führen diese Konvertierung automatisch durch. Ihr Datenmodell orientiert sich hierzu an den Konzepten, die für objektorientierte Programmiersprachen kennzeichnend sind. Hierzu gehören die Kapselung von Daten und der darauf anwendbaren Operationen (Methoden) in so genannten Objekten, die Zusammenfassung sich gleich verhaltender Objekte in Klassen sowie die Möglichkeit der Vererbung. Darauf aufbauend bieten OODBMS beispielsweise die Möglichkeit, sowohl die Struktur komplexer Objekte als auch die zugehörigen Me-

thoden zu spezifizieren. Die damit erreichbare Flexibilität kann insbesondere in Datenbankanwendungen für CAD/CAM, wissenschaftliche Experimente, Telekommunikation und Multimedia gewinnbringend genutzt werden.

XML (Extensible Markup Language) hat als Sprache für Datenaustauschformate eine zentrale Bedeutung erlangt. Es ist darum nahe liegend, die Strukturierungsmöglichkeiten, die durch XML vorgegeben sind, direkt im Datenmodell eines Datenbanksystems umzusetzen. In einem nativen *XML-Datenbankmanagementsystem (XMLDBMS)* wird dieses Ziel verfolgt. Ähnlich wie bei objektorientierten Datenbanksystemen entfällt hierdurch die Abbildung der XML-Strukturen auf ein anderes Datenmodell, wie z.B. das relationale Model. Die spezifischen Strukturierungsmöglichkeiten von XML müssen sich in diesem Fall auch in der Anfragesprache widerspiegeln. Hierfür gibt es eine Reihe von Sprachvorschlägen. Mit XQuery (siehe hierzu auch Abschnitt 4.3.3) steht hierfür eine mächtige Anfragesprache zur Verfügung, die von XML-Datenbanksystemen in der Regel unterstützt wird oder in Zukunft unterstützt werden soll.

Auch mit *multidimensionalen Datenbankmanagementsystemen* (MDBMS) wird der Ansatz verfolgt, die anwendungs-spezifische logische Strukturierung der Daten als physische Struktur zu übernehmen. Für Analyseansätze, wie beispielsweise dem Online Analytic Processing (vgl. Abschnitt 4.5.1), werden Daten häufig entlang mehrerer Dimensionen organisiert. In einem MDBMS bilden darum mehrdimensionale Felder die Basisstruktur zur Speicherung der Daten. Typische analytische Zugriffe auf die Daten werden hier besonders effizient unterstützt.

Das Datenmodell, das für einen Anwendungsbereich als besonders geeignet erscheint, direkt in einem Datenbankmanagementsystem umzusetzen, ist ein Weg, eine möglichst optimale Unterstützung der Anwendung zu erzielen. In einer realen IT-Landschaft führt diese Vorgehensweise allerdings sehr häufig zu großer Heterogenität, was die eingesetzten Datenbanksysteme angeht. Darüber hinaus bilden relationale Systeme meist den Kern der Datenhaltung eines Unternehmens. Daraus ergibt sich unmittelbar die Anforderung, relationale Datenbanksysteme so zu erweitern, dass die spezifischen Anforderungen unterschiedlicher Anwendungsbereiche berücksichtigt werden. Ein Schritt in diese Richtung wurde mit den so genannten *objektrelationalen Datenbanksystemen (ORDBMS)* beschritten. Diese bauen auf relationalen Systemen auf, integrieren aber einige objektorientierte Konzepte, die auch von objektorientierten Datenbanksystemen adressiert werden. Insbesondere gehört hierzu die Möglichkeit, benutzerdefinierte, komplexe Objekte und zugehörige Methoden zu verwalten und damit das Datenbankmanagementsystem zu erweitern. Die notwendigen Erweiterungen der Anfragesprache SQL wurden bereits 1999 in den Sprachstandard aufgenommen (vgl. Melton 2003).

Eine ähnliche Entwicklung zeichnet sich derzeit im Bezug auf XML ab. Alle kommerziellen objektrelationalen Datenbanksysteme werden in naher Zukunft mit XML-spezifischer Funktionalität ausgestattet sein, die einerseits Unterstützung für

die Abbildung zwischen XML und dem relationalen Modell bietet und andererseits eine angepasste Anfrageschnittstelle in Form einer SQL-Erweiterung bietet. Hinsichtlich dieser Erweiterung von SQL sind erste Standardisierungsschritte bereits erfolgt (vgl. ISO 2003). Es ist somit davon auszugehen, dass um XML-Funktionalität erweiterte objektrelationale Datenbanksysteme für eine Vielzahl von Anwendungen eine ausreichende Unterstützung bieten und dass ein spezielles OODBMS oder XMLDBMS nur bei sehr spezifischen Anforderungen eingesetzt werden muss.

4.4.2 Dokument Management Systeme

Dokument Management Systeme (DMS) dienen der Verwaltung von Dokumenten, die im täglichen Geschäftsablauf in Unternehmen verwendet werden. Dokumente sind Informationsträger, die inhaltlich zusammengehörende Informationen, die nicht ohne erheblichen Bedeutungsverlust weiter unterteilt werden können, zusammenfassen (vgl. Götzer et al. 2004).

Dokument Management Systeme sind durch die Weiterentwicklung von so genannten Archivsystemen entstanden. Die Kernaufgaben von Archivsystemen, die Dokumentenerfassung (Scannen), die Indizierung, die Archivierung und das Retrieval sind integrale Bestandteile heutiger DMS. Zusätzlich zur Erfassung herkömmlicher Dokumente unterstützen moderne DMS die Erstellung und Verwaltung digitaler Dokumente und ihre nahtlose Integration in unternehmensinterne Geschäftsprozesse.

Während ihrer Lebensdauer im Unternehmen durchlaufen Dokumente den so genannten Dokumentenlebenszyklus. Die einzelnen Stationen dieses Lebenszyklus werden in den entsprechenden Verarbeitungskomponenten im Dokument-Workflow (vgl. Abbildung 4-4) wiederspiegelt, das wiederum als Grundlage für die Architektur eines DMS dienen kann.

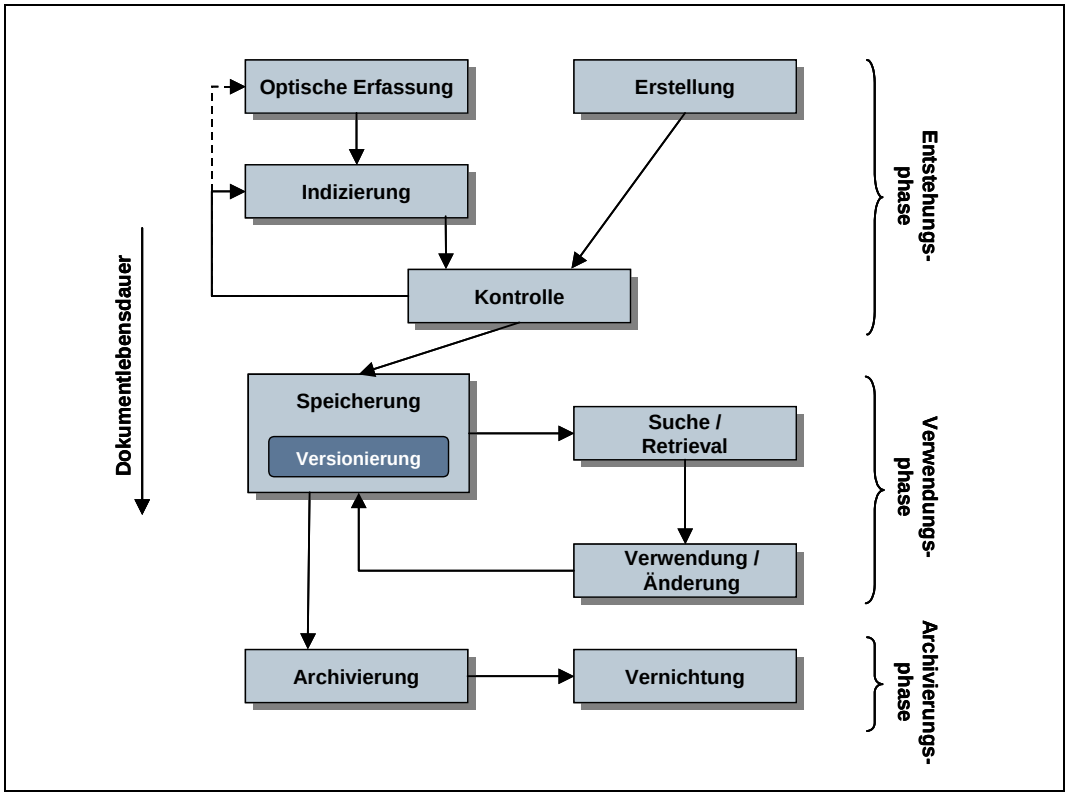

Abbildung 4-4: Dokument-Workflow (Eigene Darstellung nach Klingenhöller 2001)

Demnach gelangen Dokumente in ein DMS entweder durch die optische Erfassung gedruckter Dokumente, oder durch die direkte Erstellung in digitaler Form. Während die Ermittlung von Metadaten (z.B. Dokumenttyp, Autor, ...) für optisch erfasste Dokumente durch einen expliziten Verarbeitungsschritt (manuell oder mit OCR) geschieht, können diese Daten für digital erzeugte Dokumente automatisch erhoben werden. Die Entstehungsphase von Dokumenten wird meistens durch einen Kontrollschritt abgeschlossen. Nach der Kontrolle wird das Dokument entweder für die Speicherung und Weiterverarbeitung freigegeben, oder es muss neu indiziert oder überprüft werden.

Nach der Freigabe werden Dokumente im Speichersystem des DMS abgelegt. Optimaler Weise bietet die Speicherkomponente eine Versionsverwaltung. Nachdem Dokumente durch die Retrieval-Komponente des Systems gefunden wurden, geändert und Weiterverarbeitet wurden, sollen diese in einer geänderten Version wieder im Speichersystem abgelegt werden.

Nach der aktiven Verarbeitungsphase werden Dokumente im Langzeitarchiv des Systems abgelegt. Während die Speicher-Komponente des DMS auf hohe Verarbeitungsgeschwindigkeit (Hauptspeicher, RAID-Systeme) ausgelegt ist, können für

die Langzeitarchivierung günstigere Speichermedien (z.B. Bandarchive) eingesetzt werden.

Einflussfaktoren, die die zukünftige Entwicklung von DMS und ihren Einsatz in Unternehmen bestimmen, sind zum Beispiel die Menge der zu verwaltenden Dokumente oder die zunehmende Mobilität, die von Mitarbeitern verlagt wird. Dokumente, die in täglichen Geschäftsprozessen gebraucht werden, müssen auch unterwegs jederzeit verfügbar sein. Daher wird die Entwicklung von Unternehmensportalen angestrebt, die gestützt durch DMS die ortsunabhängige Verwendung von Dokumenten erlauben. Aus diesem Grund werden DMS schon heute mit Content-Management-Funktionalität ergänzt (vgl. Kap. 4.6.1).

4.4.3 Digitale Bibliotheken

Im Gegensatz zu DMS liegt der Schwerpunkt von Digitalen Bibliotheken auf der Speicherung von Dokumenten, die bereits in digitaler Form vorliegen. Ein weiterer Aspekt, der Digitale Bibliotheken auszeichnet, ist die Verfügbarkeit über das WWW. Ein prominentes Beispiel einer digitalen Bibliothek ist das *Networked Computer Science Technical Research Library (NCSTRL)*, dessen Hauptkomponenten in Abbildung 4-5 dargestellt werden.

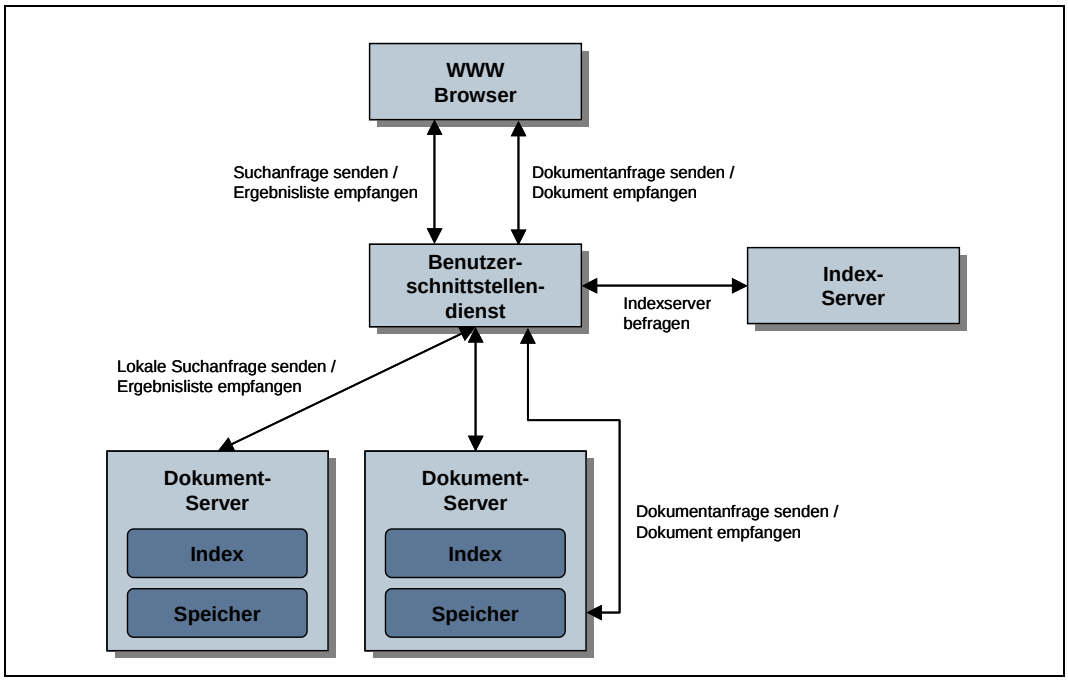

Abbildung 4-5: NCSTRL Dienste (Eigene Darstellung nach Davis 2000)

Wie der Name schon sagt, ist NCSTRL eine über das Internet verteilte Digitale Bibliothek für die Verwaltung von wissenschaftlichen Veröffentlichungen. Über die

Webschnittstelle des Systems können Suchanfragen gestellt und Dokumente heruntergeladen werden. Der Benutzerschnittstellendienst befragt den Index-Server und liefert Ergebnislisten für die Benutzersuchanfragen. Digitale Dokumente werden direkt von den im Internet verteilten Dokument-Servern geholt.

Zusammenfassend kann man sagen, das der Einsatz einer Digitalen Bibliothek erst dann Sinn macht, wenn viele digital erstellte Dokumente, die hinterher nicht geändert werden anfallen.

4.5 Technologien für die Analyse

Wie bereits oben erwähnt existieren vielfältigste Möglichkeiten, Daten zu analysieren. In diesem Abschnitt werden mit Online Analytic Processing und Data Mining zwei wichtige Technologien vorgestellt, die es erlauben, neue Erkenntnisse aus umfangreichen Datenbeständen zu gewinnen.

Beide Analyseansätze setzen strukturierte Daten voraus. Diese werden in der Regel in einer Data-Warehouse-Datenbank bereitgestellt. Als Speicherungstechnologie sind hierbei relationale und objekt-relationale Datenbanksysteme von zentraler Bedeutung. Die im Rahmen von OLAP-Analysen bereitgestellten, aggregierten Daten werden häufig in multidimensionalen Datenbanksystemen verwaltet (vgl. Abschnitt 4.4.1). Die Daten, die in der Data-Warehouse-Datenbank integriert werden, stammen in der Regel zu einem großen Teil aus der Organisation, in der die Analyse durchgeführt wird, und sind dort auf mehrere Datenquellen verteilt. Hierbei kann es sich beispielsweise um die Datenbanken unterschiedlicher Fachabteilungen handeln. Bevor diese heterogenen Datenbestände konsolidiert in der Data-Warehouse-Datenbank integriert werden, findet eine Verknüpfung mit zusätzlichen externen Datenbeständen statt, soweit dies für die Analysen erforderlich ist. Nichtsdestotrotz liegt insbesondere bei OLAP der Fokus auf der Verarbeitung unternehmensinterner, strukturierter Daten.

4.5.1 Online Analytic Processing

Online Analytic Processing (OLAP) ermöglicht die mehrdimensionale Analyse von Daten, die in einer Data-Warehouse-Datenbank vorliegen sowie die meist grafische Darstellung der Analyseergebnisse (vgl. Dinter et al. 1998). Die Dimensionen entsprechen hierbei den unterschiedlichen Blickwinkeln, aus denen die Anwender vorhandene Daten typischerweise analysieren, wie z.B. die Zeit oder geographische Bezüge. Die verschiedenen Dimensionen weisen wiederum eine interne Struktur auf. Sie umfassen mehrere, hierarchisch strukturierte Attribute. Beispielsweise besteht eine Zeitdimension häufig aus den Attributen Tag, Woche, Monat und Jahr. Die für die Analyse zentralen Daten werden als Fakten bezeichnet, die auf die einzelnen Dimensionen bezogen vorliegen. Durch die Betrachtung der Fakten entlang verschiedener Dimensionen ergibt sich im dreidimensionalen Fall die Struktur eines Datenwürfels. Unabhängig von der physischen Speicherungsstruk-

tur wird die Metapher eines Datenwürfels (engl. Data Cube) für die Verdeutlichung der typischen OLAP-Operationen verwendet. Wichtige Operationen werden im Folgenden kurz am Beispiel eines Datenwürfels, der Umsatzzahlen bezogen auf die Dimensionen Zeit, Produkt und Filiale enthält, erläutert (siehe Abbildung 4-6 (a)).

Mit einer Selektion wird die Analyse auf einzelne Attributwerte innerhalb einer Dimension fokussiert. Bezieht sich die Einschränkung auf eine einzige Dimension, so wird durch die Operation (engl. *slice*) eine Ebene des Datenwürfels herausgeschnitten. In Abbildung 4-6 (b) sind dies die Umsatzzahlen für den Monat 12/04. Wird die Analyse in zwei oder mehr Dimensionen auf einzelne Werte eingeschränkt, so ist das Ergebnis ein Teilwürfel der ursprünglichen Daten. In Abbildung 4-6 (c) ist der Teilwürfel für den Monat 12/04, das Produkt MonitorX7 und die Filiale München hervorgehoben.

Die Operation *drill-down* ermöglicht eine Navigation auf den Daten, indem sie zu dem betrachteten Datenwürfel zusätzliche Detailinformation liefert. Dies kann einerseits durch das Absteigen innerhalb der Hierarchie einer Dimension erfolgen. Andererseits können die Daten auch bezüglich einer oder mehrerer zusätzlicher Dimensionen aufgeschlüsselt werden. Die gegenteilige Operation, als *drill-up* oder *roll-up* bezeichnet, ermöglicht den Wechsel auf eine höhere Hierarchiestufe innerhalb einer Dimension bzw. das Eliminieren einer Dimension des Datenwürfels. Beides ist mit einer Aggregation der Daten verbunden. Drill-down ist in Abbildung 4-6 (e) und drill-up in Abbildung 4-6 (f) dargestellt. Ausgangspunkt ist jeweils der Datenwürfel aus Abbildung 4-6 (d). In diesem Fall bedeutet drill-down, dass die zunächst auf Monatsebene verfügbaren Daten für die einzelnen Tage aufgeschlüsselt werden. Beim drill-up findet dagegen eine zusätzliche Aggregation statt, so dass die Umsätze pro Jahr verfügbar gemacht werden.

Eine weitere Operation, die sich auf die Darstellung der Daten bezieht, ist *pivot*. Diese ermöglicht den Wechsel der Achsen eines Datenwürfels, so dass dieser mit einer anderen Orientierung im Raum dargestellt wird. Dies ermöglicht die Betrachtung der Daten aus unterschiedlichen Blickwinkeln und ist insbesondere dann relevant, wenn mehr als drei Dimensionen zu berücksichtigen sind.

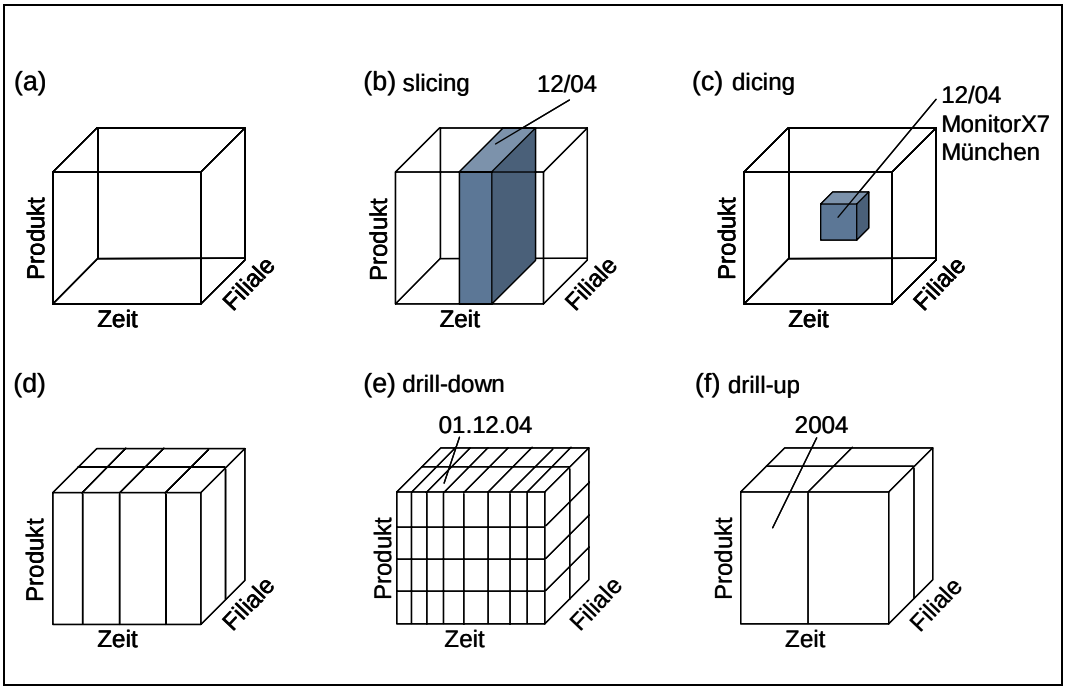

Abbildung 4-6: Typische OLAP-Operationen

OLAP-Werkzeuge bieten in der Regel unterschiedlich komplexe Schnittstellen, die sich je nach Nutzergruppe zuschneiden lassen. In einer einfachen Variante erlaubt ein Werkzeug das Aufrufen bereits vorbereiteter Analysen, für die eventuell noch einzelne Parameter angegeben werden können. Die komplexeste Schnittstelle ermöglicht es den Nutzern dagegen, auf alle in einem Data Warehouse verfügbaren Daten zuzugreifen und hierzu beliebige Kombinationen der oben genannten Operationen einzusetzen. Diese Schnittstelle erlaubt es, individuelle Analysen ad-hoc zu definieren.

4.5.2 Data Mining

Ein weiterer wichtiger Analyseansatz für große Datenbestände wird als *Data Mining* bezeichnet. Mit diesem Ansatz wird das Ziel verfolgt, Muster und Zusammenhänge in Daten zu erkennen, die so vorab nicht bekannt sind und die für eine bestimmte Anwendung sinnvoll eingesetzt werden können. Hierbei müssen die Anwender ihren Informationsbedarf wesentlich weniger genau spezifizieren können, als dies bei anderen Analyseansätzen, wie z.B. OLAP der Fall ist. OLAP unterstützt die Überprüfung von Hypothesen, die durch den Anwender aufgestellt wurden. Im Gegensatz dazu sollen Data-Mining-Werkzeuge möglichst selbständig neue Hypothesen aufstellen.

Mit dem Begriff Data Mining wird eine ganze Reihe unterschiedlicher Ansätze verbunden (vgl. Han und Kamber 2001). Hierbei muss zwischen dem zu

erstellenden *Data-Mining-Modell*, der hierfür gewählten *Repräsentation*, den Kriterien zur Bewertung von Modellen und den *Data-Mining-Algorithmen* zur Ableitung der Modelle unterschieden werden. Es gibt unterschiedliche Arten von Data-Mining-Modellen. Jedes kann bestimmte Muster und Zusammenhänge zu gegebenen Daten darstellen.

Unter *Clustermodellen* versteht man die Unterteilung einer Menge von Objekten in mehrere Klassen. Hierbei sollen sich die Objekte, die derselben Klasse zugeordnet werden, möglichst ähnlich sein, während sie sich in ihren Merkmalen von den Objekten anderer Klassen möglichst stark unterscheiden sollen. Eine mögliche Repräsentation für ein Clustermodell beinhaltet für jedes gefundene Cluster die normierten Attributwerte des Objekts im Zentrum des Clusters. In Abbildung 4-7 (a) ist diese Repräsentation von Clustern durch die zentralen Objekte dargestellt.

Bei den *Klassifikationsmodellen* werden dagegen die Regeln identifiziert, nach denen Objekte vorgegebenen Klassen zugeordnet werden können. Eine wichtige Repräsentationsform in diesem Bereich sind *Entscheidungsbäume*. Ein einfaches Beispiel für einen Entscheidungsbaum zeigt Abbildung 4-7 (b). Hier wird dargestellt, wie Kunden auf Grund der Eigenschaften Alter, Beruf und Kreditwürdigkeit einer von zwei Klassen zugeordnet werden können. Die beiden Klassen umfassen Kunden, die einen Computer kaufen (Knoten mit der Bezeichnung ‘ja’) und solche die dies nicht tun (Knoten mit der Bezeichnung ‘nein’). Alternativ können Klassifikationsmodelle auch als Wenn-Dann-Regeln, mathematische Formeln oder mit Hilfe neuronaler Netze dargestellt werden.

Einen Ansatz, um zu beschreiben, wie aus gegebenen Eigenschaften von Objekten andere, kontinuierliche Werte abgeleitet werden können stellen *Regressionsmodelle* dar. Diese Modelle werden durch lineare oder nicht-lineare Gleichungen repräsentiert.

Abhängigkeitsmodelle beschreiben signifikante Zusammenhänge sowohl zwischen Attributen von Objekten als auch zwischen Daten in einer Datenbank. Letztere können z.B. durch *Assoziationsregeln* repräsentiert werden. Eine textbasierte und eine grafisch unterstützte Möglichkeit solche Assoziationsregeln anzugeben ist in Abbildung 4-7 (c) dargestellt. Die erste Regel beschreibt beispielsweise, dass 75% aller Kunden, die sowohl einen Computer als auch einen Monitor kaufen, bei demselben Einkauf auch noch einen Drucker erwerben.

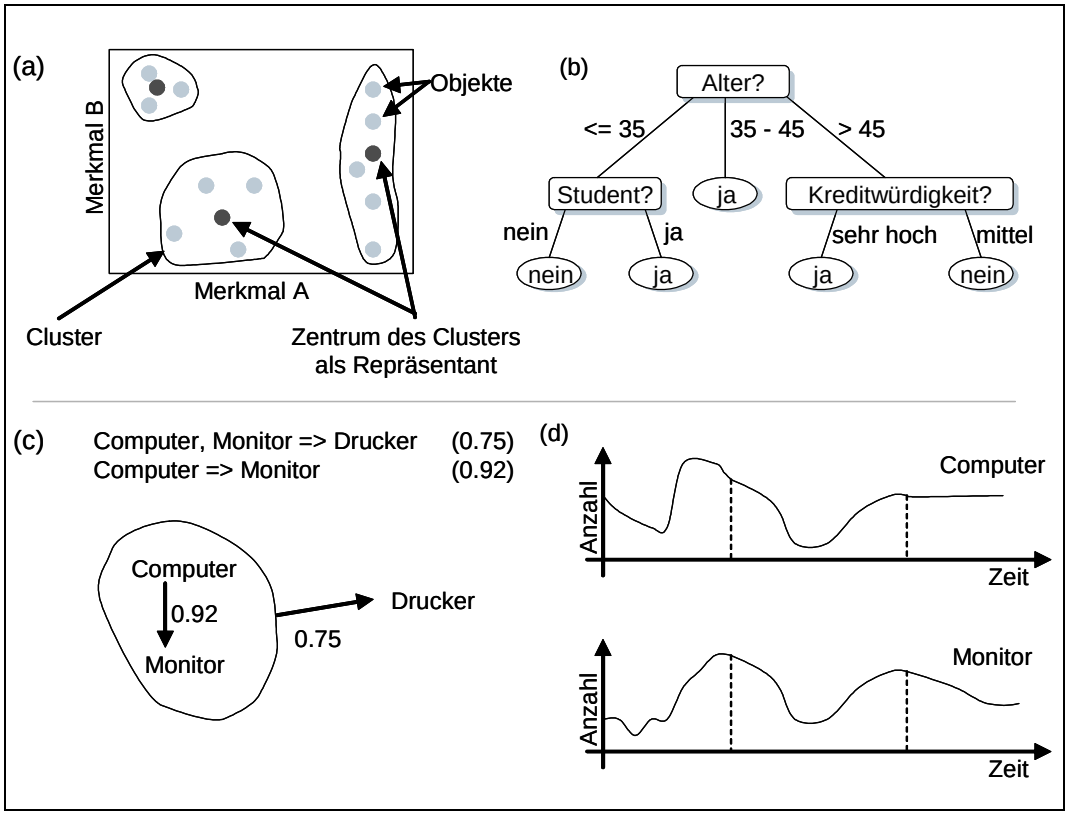

Abbildung 4-7: Data-Mining-Modelle

*Z*eitabhängige Muster in Datenströmen werden durch *Sequenzmodelle* beschrieben. Hier ist es das Ziel, den Prozess, der die Daten generiert, zu beschreiben bzw. Abweichungen, Muster und Trends in einer Sequenz zu erkennen. Abhängig von den Ausgangsdaten können Sequenzmodelle durch Gleichungen oder durch geordnete Listen von Objekten repräsentiert werden. Ein Beispiel für zwei Sequenzen, hier die Verkaufszahlen für Computer und Monitore, ist in Abbildung 4-7 (d) gegeben. Für beide Sequenzen wurde ein gemeinsames Muster identifiziert.

Neben den unterschiedlichen Modellen und deren Repräsentationsmöglichkeiten müssen beim Data Mining die Kriterien festgelegt werden, nach denen unterschiedliche Modelle bewertet und verglichen werden können. Diese Kriterien werden teilweise von den Algorithmen zur Generierung der Modelle genutzt. Es existiert ein breites Spektrum an Algorithmen, von denen jeder nur ein spezifisches Modell bzw. eine bestimmte Repräsentationsform abdeckt. Für jedes Modell oder jede einzelne Repräsentationsform kann es aber wiederum mehrere Algorithmen geben, denen in der Regel statistische Verfahren zu Grunde liegen. Eine Übersicht verfügbarer Algorithmen gibt es z.B. in Han und Kamber 2001.

Die Gesamtheit aller Data-Mining-Algorithmen, die geeignet sind, eine spezifische Art von Data-Mining-Modell abzuleiten, kann als *Data-Mining-Verfahren* bezeichnet werden. Somit stellen die *Clusterbildung*, die *Klassifikation*, die *Regressionsanalyse*, die *Abhängigkeitsanalyse* und die *Sequenzanalyse* die zu den oben aufgeführten Modellen zugehörigen Data-Mining-Verfahren dar.

Data-Mining-Verfahren werden den Anwendern in unterschiedlicher Form zur Verfügung gestellt. Einerseits exisiteren spezielle *Data-Mining-Werkzeuge*, die in der Regel mehrere Data-Mining-Verfahren anbieten und zusätzlich Funktionalität zur Vorverarbeitung der Daten sowie zur grafischen Aufbereitung der abgeleiteten Modelle anbieten. Andererseits bilden Data-Mining-Werkzeuge auch einen integralen Bestandteil anderer Anwendungen, wie z.B. Customer-Relationship-Management-Systemen.

4.6 Technologien für die Bereitstellung

4.6.1 Content Management Systeme

Wie in Kapitel 3 bereits dargelegt, sind Daten, Information und Wissen im Kontext des Innovationsmanagements Faktoren, die nicht nur für den Erfolg oder Misserfolg eines einzelnen Innovationsprojektes, sondern auch für den langfristigen Erfolg des gesamten Unternehmens von entscheidender Bedeutung sind. Eine Studie des Fraunhofer IAO (vgl. Wilhelm 2003) aus dem Jahr 2003 belegt, dass Content Management Systeme in der Produktion erfolgreich eingesetzt werden können. Das benötigte Wissen sollte jedem Mitarbeiter zur richtigen Zeit in geeigneter Darstellungsform zur Verfügung stehen. Gerade in großen, multikulturellen Unternehmen ist diese Aufgabe ohne geeignete, informationstechnische Unterstützung nicht realisierbar. Content-Management-Systeme (CMS) und insbesondere Web-Content-Management-Systeme (WCMS), die die systematisierte Verbreitung von Inhalten im Internet und in Intranets ermöglichen, schaffen hierbei Abhilfe.

Zentral für CMS sind die Inhalte (*Content*), die publiziert werden. Unter dem Begriff *Inhalt* verbergen sich Texte, Bilder, Videos und noch eine Reihe von weiteren Datendarstellungsformen. Im Bereich des Content Managements werden diese oft als *Assets* (vgl. Rothfuss 2003) bezeichnet.

Der Schwerpunkt eines CMS liegt in der Regel bei der Verwaltung und Publikation von Inhalten. Ein weiterer, wichtiger Aspekt kann die Contenterstellung sein, weswegen solche Systeme auch oft als *Redaktionssysteme* bezeichnet werden.

Während der Bearbeitung in einem CMS durchlaufen Inhalte einen Lebenszyklus (*Content Life Cycle*), der in Abbildung 4-8 dargestellt wird. Erstellte Inhalte werden in der Regel durch eine andere Instanz kontrolliert. Nach einer eventuellen Überarbeitung folgt die Freigabe und schließlich die Publikation. Am Ende des Inhaltslebenszyklus steht die Archivierung.

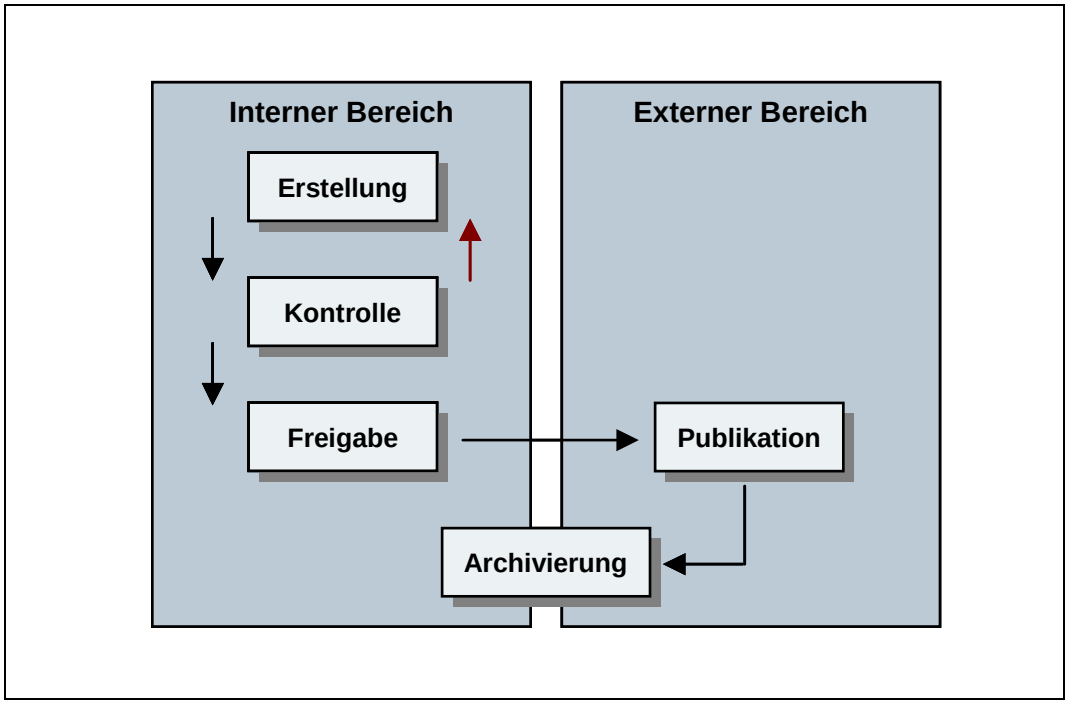

Abbildung 4-8: Der Content Life Cycle (Quelle: Zschau et al. 2002)

Die einzelnen Phasen der Content-Verarbeitung werden in der Regel auf verschiedene Arbeitsplätze aufgeteilt. Ein ideales CMS muss aus diesem Grund den Inhaltsfluss (Content Workflow) durch einen modularisierten Aufbau, wie in Abbildung 4-9 dargestellt, unterstützen. Ein besonderes Augenmerk muss beim Content Management auf die Trennung von *Inhalt* und *Darstellung* gelegt werden. Nur wenn dies gewährleistet ist, können Inhalte auf einfache Art verschiedenen Benutzergruppen in unterschiedlicher Darstellung zugänglich gemacht werden. In der gezeigten Architektur schlägt sich dieses Grundkonzept in dem getrennten Speichermodul (Content Repository) und dem Darstellungsmodul (Content-Präsentation) nieder.

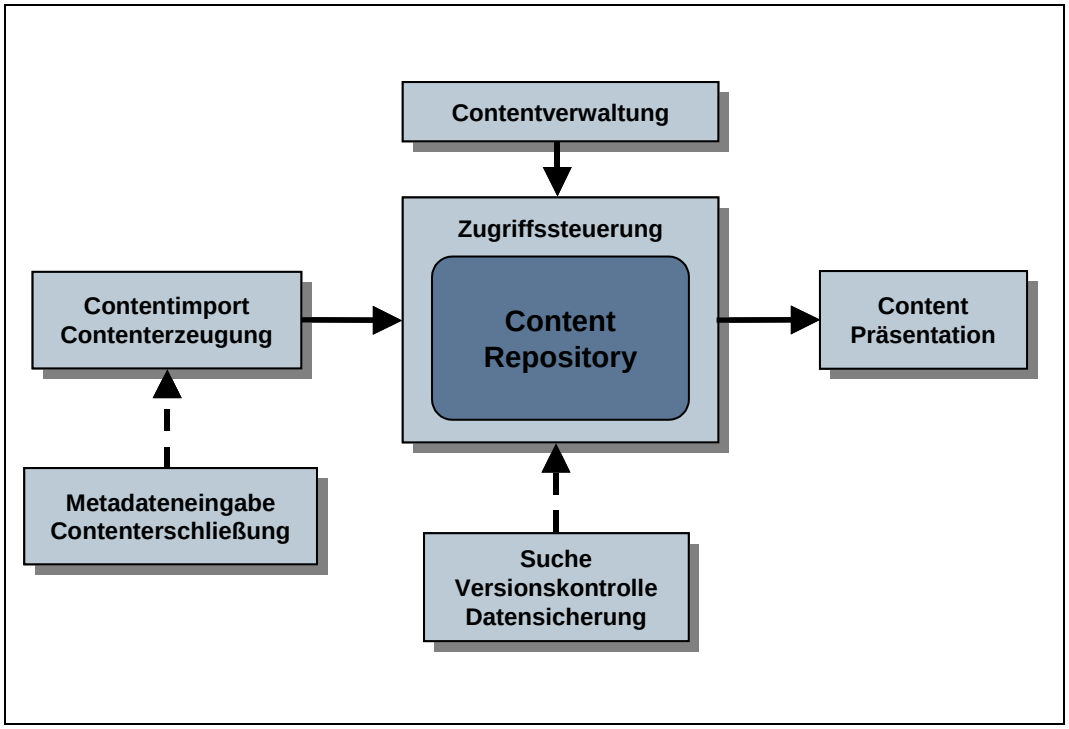

Abbildung 4-9: Architektur eines Content Management Systems (Eigene Darstellung nach Rothfuss 2003)

Für die kommenden Jahre ist zu erwarten, dass es in Bereich der CMS zahlreiche Weiterentwicklungen geben wird. Zum einen haben Unternehmen erkannt, dass die Publikation von Inhalten (sowohl nach außen, als auch innerhalb des Unternehmens) mit einem CMS leichter zu bewältigen ist, zum anderen besitzen heute existierende CMS erhebliches Verbesserungspotenzial.

Eine der Entwicklungsrichtungen, die bereits heute zu erkennen ist, geht in Richtung der Unterstützung von Personalisierung und Portalfunktionen für CMS. Ein anderer Aspekt, der auf die direkte Verwertung von Inhalten abzielt, ist die Verteilung/Verkauf von Inhalten (Content Syndication) zwischen verschiedenen Unternehmen und deren CMS.

Aus technologischer Sicht ist es wahrscheinlich, dass XML eine verstärkte Rolle im Bereich des Content Managements spielen wird. Die Forderung nach der Trennung von *Inhalt* und *Darstellung* kann mit XML-Technologien, wie XML-Schema, XSLT und XQuery hervorragend erfüllt werden.

4.6.2 Präsentation

Die Präsentation von gespeicherten Informationen erfolgt zunehmend über Schnittstellen und Werkzeuge, die auch in den übrigen Bereichen der Datenverar-

beitung erfolgreich eingesetzt werden. Allen voran ist hier die Integration in Webserver- und Browser-Umgebungen zu nennen. Dadurch bietet sich dem Benutzer für immer mehr Aufgaben eine einheitliche Schnittstelle, was den Umgang mit unterschiedlichen Systemen deutlich erleichtert und die Informationsverarbeitung somit effizienter macht.
Auch in diesem Zusammenhang spielt XML eine immer stärkere Rolle. Wie bereits oben angeführt, sollte möglichst darauf geachtet werden, die eigentlichen Informationen von ihrer Darstellung, d.h. vom Layout getrennt zu halten. Nur so ist es ohne großen Aufwand möglich, verschiedene Sichten auf eine Informationsbasis zu bieten und unterschiedliche Anwendungen und Ausgabeformate zu integrieren. Wird XML als Speicherformat eingesetzt bzw. lässt sich aus dem tatsächlichen Speicherformat eine adäquate XML-Repräsentation generieren, so können diese Daten mittels Transformationen in eine Vielzahl von im jeweiligen Kontext benötigten Präsentationsformaten überführt werden (vgl. Abschnitt 4.3 sowie Abbildung 4-10).

Im o.g. Browserkontext ist dieses Ausgabeformat in der Regel serverseitig generiertes HTML. Eine Verlagerung der Transformation vom Server zum Client wird verschiedentlich vorgeschlagen, um Server-Ressourcen zu sparen und die Flexibilität der clientseitigen Darstellung zu erhöhen. Nachteilig an diesem Verfahren ist allerdings, dass entsprechende Transformationssoftware damit auf jedem Client zur Verfügung stehen muss. Gerade bei mobilen Endgeräten ist dies heute aufgrund mangelnder Bandbreite der Verbindung sowie geringer Speicher- und Rechenkapazitäten oft nicht gegeben.

Das Erzeugen des Layouts der vorstrukturierten Serverdaten erfolgt dagegen meist auf Clientseite (vgl. Wöhr 2004). Dies geschieht im Falle von HTML mit *Cascading Stylesheets (CSS)*. Dabei werden die layoutlosen Daten mit Formatierungsvorschriften angereichert und ergeben in dieser Kombination ein formatiertes HTML-Dokument. Der Einsatz von CSS bringt hier zwei Vorteile mit sich: zum einen kann der Aufwand für die Wartung der zu präsentierenden Daten stark reduziert werden, da nahezu alle Fragen des Layouts für alle Dokumente zentral in einem bzw. wenigen Stylesheets geregelt werden können. Zum anderen wird die Weiterverarbeitung durch andere Programme dahingehend erleichtert, dass sich Quelldokumente im Wesentlichen auf ihre tatsächlichen Informationen reduzieren lassen und nicht durch zusätzliche, für die maschinelle Informationsverarbeitung irrelevante Layoutdaten aufgebläht werden.

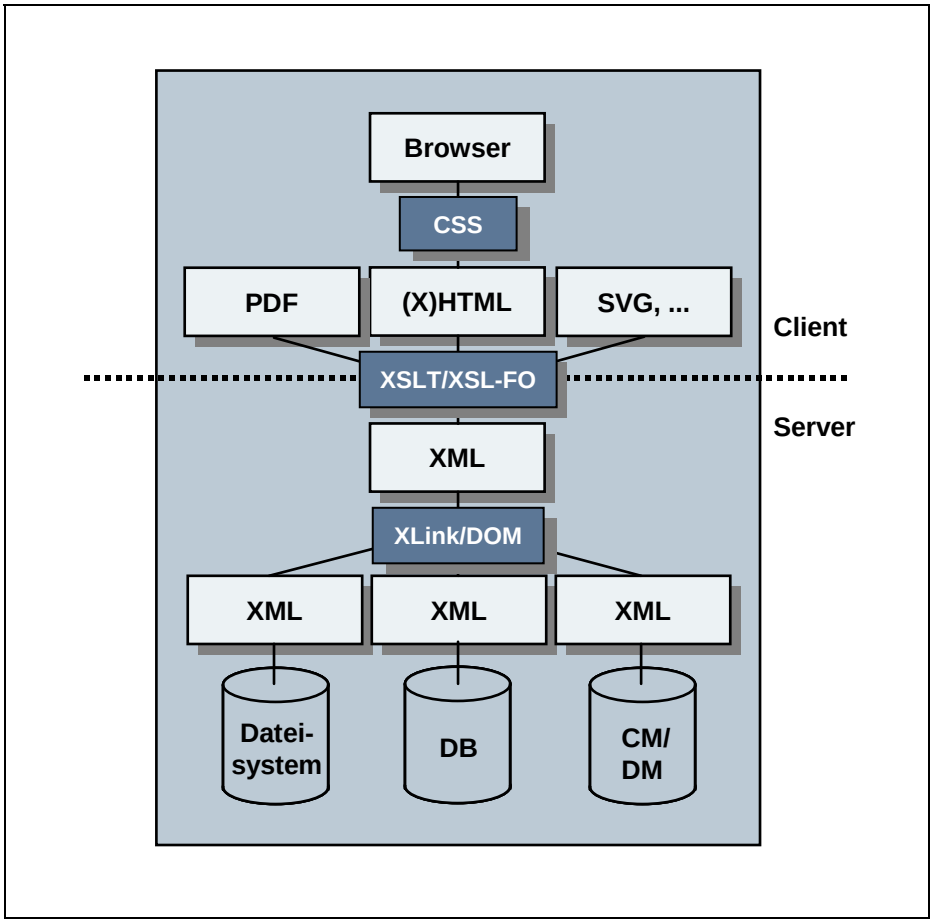

Abbildung 4-10: Einsatz verschiedener Datenformate

Ein weiterer Schritt in diese Richtung ist der zunehmende Einsatz von *XHTML* anstelle von HTML. XHTML ist prinzipiell ähnlich zu HTML, baut aber auf XML auf. Der Vorteil von XHTML gegenüber HTML liegt darin, dass ein Dokument leicht mittels XML Parsern auf seine syntaktische Korrektheit geprüft werden kann und seine logische Struktur sich zwingend aus der syntaktischen Struktur ergibt. Diese wiederum ist Voraussetzung für ein einfaches automatisiertes Verarbeiten der Daten. Die syntaktische Korrektheit des Dokuments stellt bei HTML jedoch oft ein Problem dar, da die meisten Browser verschiedene Abweichungen von der Standardsyntax unterstützen und in der Folge immer mehr „fehlerhafte“ HTML-Dokumente entstanden sind. Diese nicht standardkonformen Dokumente können von Browsern dennoch mehr oder weniger korrekt angezeigt werden, eignen sich aber für eine maschinelle Weiterverarbeitung außerhalb des Browsers nur bedingt.

Erweitert wird HTML zusätzlich durch den *XForms*-Standard, der eine ebenfalls auf XML basierende Technik bietet, Benutzereingaben aufzunehmen und an einen Server weiterzuleiten. Anders als im herkömmlichen HTML-Standard gestaltet sich

XForms flexibler in der Unterstützung unterschiedlicher (auch mobiler) Endgeräte und bietet die Möglichkeit, Benutzereingaben ohne Skript-Unterstützung bereits clientseitig zu prüfen, bevor diese an den Server geleitet werden. Dies spart Bandbreite und Serverrechenzeit, vor allem erhält der Benutzer aber eine direkte Rückmeldung über fehlerhafte oder fehlende Eingaben.

Ein weitere Neuerung im Bereich Benutzerinteraktion stellt die Autorensprache *Synchronized Multimedia Integration Language (SMIL)* dar, mit deren Hilfe sich interaktive Präsentationen im Rahmen von XML beschreiben und ausführen lassen (vgl. W3C 2001c). Neben der rein visuellen Präsentation wird zudem Unterstützung für Audiodaten angeboten, was den multimedialen Charakter unterstreicht. Dem Autor bieten sich mit SMIL verschiedene Möglichkeiten, Inhalte zeitlich synchronisiert zu präsentieren, wobei verschiedene Funktionen für Layout, Animation und Seitenübergänge sowie das einbetten externer Objekte zur Verfügung stehen. Durch die Integration von RDF existiert zudem die Möglichkeit, maschinenverständliche Metadaten in die Präsentation aufzunehmen. Da für solche Aufgaben bisher i.d.R. das Flash-Format oder vollständig vom Web losgelöst PowerPoint o.ä. eingesetzt wurde, konnte in diesen Präsentationen enthaltenes Wissen kaum automatisiert weiterverwendet werden. Es bleibt abzuwarten, ob SMIL sich gegenüber diesen Techniken wird durchsetzen können.

4.6.3 Webserver Technologien

Wie oben bereits angeführt, werden Anwendungen zunehmend unter einer gemeinsamen Web-Oberfläche integriert, um dem Benutzer einen einheitlichen Zugang zu unterschiedlichen Systemen zu bieten. Dies erfordert auch auf Seiten des Servers eine weitergehende Flexibilisierung und standardisierte Schnittstellen, um mit den zugrunde liegenden externen Anwendungen interagieren zu können. Um dies zu erreichen werden in modernen Webservern wie *Apache* oder *Microsoft IIS* Schnittstellen zu Erweiterungsmodulen angeboten. Diese lassen sich unterscheiden nach der Art der Schnittstelle, bzw. auf welche Art und Weise externe (i.d.R. dynamische) Inhalte in die Antwort des Servers an den Client einfließen können. Eine Möglichkeit besteht in der Erweiterung des statischen Inhalts von Webseiten durch in diese eingebettete Programmlogik. Dabei wird die angeforderte Datei vor ihrer tatsächlichen Auslieferung von einem Servermodul gelesen und die darin enthaltenen Programmanweisungen ausgeführt. Das an den Client zurückgelieferte Resultat wird dann durch den statischen Inhalt zusammen mit dem auf diese Art und Weise generierten Dynamischen Inhalt gebildet. Diese Methode eignet sich speziell für Anwendungen, in denen lediglich einzelne Abschnitte einer Seite dynamisch generiert werden müssen und wird durch Sprachen/Techniken wie PHP oder allgemein *Server Side Includes (SSI)* realisiert.

Im Gegensatz zum dokumentenorientierten Ansatz von PHP und SSI sind Skriptsprachen wie Perl oder Python stärker programmorientiert und werden vom Webserver über das *Common Gateway Interface (CGI)* oder als Server-Modul einge-

bunden. Dabei wird die Erstellung des gesamten auszuliefernden Inhalts dem Skript überlassen, was eine größere Flexibilität hinsichtlich der Gestaltung jeder einzelnen Seite mit sich bringt.

Den mächtigeren Ansatz stellen jedoch *Java Servlets* und *Java Server Pages (JSP)* dar. Hier steht der volle Funktionsumfang der Sprache Java zur Verfügung und zudem kann leicht eine vollständige Trennung von Programmlogik und Präsentation erreicht werden. Servlets liegen als kompilierte Module vor und werden durch einen Servletcontainer (z.B. *Tomcat*) verwaltet und ausgeführt. Das spart den Aufwand für das Starten von externen Prozessen (vgl. beispielsweise CGI) und bringt zusätzliche Erleichterungen und Performance-Verbesserungen dadurch, dass der Container Sitzungsdaten und evtl. Datenbankverbindungen über einen längeren Zeitraum hinweg verwalten kann, ohne sie bei jedem Client-Zugriff neu initialisieren zu müssen.

Ebenso seien an dieser Stelle *Active Server Pages (ASP)* erwähnt. Sie kombinieren Eigenschaften von dokumenten- und programmorientierten Techniken. Zwar werden ASP über Skriptelemente (JSkript oder VBSkript) interpretiert (vgl. PHP), sie bieten jedoch über die Windows (D)COM-Infrastruktur Zugriff auf eine beliebig mächtige Funktions- und Servicebibliothek, ähnlich dem Ansatz mit Java Servlets. Mit ASP.NET (vgl. 4.7.4) wird die Interpretation zugunsten von kompiliertem Code aufgegeben. Nachteilig ist dennoch, dass der Einsatz von ASP auf Server mit Windows-Betriebssystemen beschränkt ist.

Die genannten Methoden lassen sich alle mit den im vorigen Abschnitt angeführten Techniken für Inhaltsaufbereitung und Layout (XML, XSLT, CSS, etc.) verbinden. Somit entstehen aus ihrer Kombination mächtige Werkzeuge und es ergeben sich vielfältige Einsatzmöglichkeiten, um Inhalte dynamisch, dem Kontext entsprechend darstellen zu können (vgl. Wöhr 2004).

Ein weiterer Aspekt der Informationspräsentation wird mit dem W3C-Projekt *annotea* im Rahmen des Semantic Web abgedeckt (vgl. W3C 2001b). Annotea ermöglicht es dem Benutzer, Notizen und Lesezeichen zu einzelnen Webressourcen bzw. beliebigen XML-Dokumenten zu erzeugen und zu verwalten. Realisiert wird dies auf Basis von Browsererweiterungen auf Seiten des Clients und eigenständigen sog. Annotation-Servern, die entsprechende Daten personen- und ressourcenbezogen verwalten. Die Stärke dieser Vorgehensweise liegt darin, dass zentral verwaltete Annotationen von beliebigen Clients aus genutzt werden können (vgl. im Gegensatz dazu das Bookmarksystem heutiger Browser) und zudem auf einfache Art und Weise anderen zugänglich gemacht werden können.

4.7 Querschnittstechnologien

Mehr und mehr Anwendungen werden mittlerweile nicht mehr zentral auf einem Rechner ausgeführt, sondern auf mehre Systeme verteilt. Für den Benutzer ge-

schieht dies jedoch meist transparent, d.h. komplexe Systeme aus verteilten Komponenten präsentieren sich ihm mit einer einheitlichen Schnittstelle, hinter der die Interna der tatsächlichen Realisierung verborgen bleiben. Die Verteilung selbst kann wiederum auf mehreren Ebenen geschehen (vgl. Abbildung 4-11).
Die im Folgenden beschriebenen Technologien lassen sich schwer genau einer der im gesamten Kapitel 4 beschriebenen Phasen des Information Workflow zuordnen, vielmehr bauen sie auf diesen auf bzw. diese auf ihnen. Durch ihre Verknüpfung bilden sie im Kontext dieser Übersicht Querschnittstechnologien.

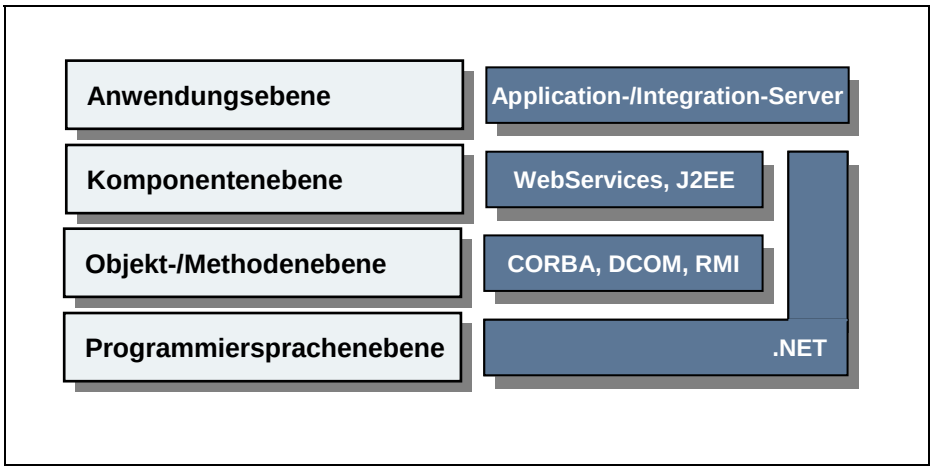

Abbildung 4-11: Ebenen der Systemvernetzung mit aktuellen Beispielen

4.7.1 Interoperabilität von Programmiersprachen

Bei der Programmierung von netzbasierten Anwendungen kommt potenziell eine Vielzahl an Programmiersprachen zum Einsatz. Da für solche Anwendungen ein relativ breit gefächertes Technologiefeld genutzt wird, kommen hier die Unterschiede der Programmiersprachen deutlich zum Tragen, da sich unterschiedliche Sprachen für unterschiedliche Zwecke besonders eigenen. Für systemnahe Funktionalität werden z.B. meist Sprachen wie C oder C++ eingesetzt, wohingegen zur Programmierung von graphischen Oberflächen eher VisualBasic oder Skriptsprachen genutzt werden. Letztere haben zwar den Nachteil einer geringeren Performanz, bieten jedoch einige Konzepte, die es dem Programmierer erleichtern, sichere und fehlerarme Anwendungen zu schreiben.
Idealerweise würden zur Entwicklung einer Anwendung mehrere Sprachen eingesetzt, eine jede für den von ihr besonders unterstützten Teilbereich. Dies gestaltet sich jedoch in der Regel kompliziert, was sich hauptsächlich auf inkompatible Typensysteme der einzelnen Programmiersprachen zurückführen lässt.
Dieser Problematik begegnet die von Microsoft entwickelte .NET-Technologie (vgl. Beer 2003 sowie Abschnitt 4.7.4). Sie ermöglicht es, beliebige Teile einer Anwendung in unterschiedlichen (.NET-fähigen) Sprachen zu entwickeln und die einzelnen Komponenten zu einer einzigen Anwendung zu integrieren. Realisiert wird

dies dadurch, dass, im Gegensatz zum herkömmlichen Compilier-Vorgang, der Compiler keinen Maschinencode erzeugt, sondern einen sog. Intermediate-Code, also einen nicht direkt ausführbaren Zwischencode. Dieser Zwischencode fungiert als eine gemeinsame Schnittstelle zwischen den o.g. Programmiersprachen und dient unter anderem der Spezifikation verwendeter Variablentypen. Um das Programm auszuführen, wird dieser Zwischencode beim Start von einer Laufzeitumgebung in Maschinencode übersetzt und daraufhin ausgeführt. Hier zeigt sich auch ein wesentlicher Unterschied zur Sprache Java. Java-Code wird ebenfalls in einen Zwischencode übersetz, allerdings wird dieser nicht notwendigerweise zur Laufzeit in Maschinencode umgewandelt, sondern in der Regel von einer sog. Virtual Machine interpretiert. Die .NET-Lösung verspricht hier eine etwas höhere Performanz.

4.7.2 Verteilte Ausführung eines Programms über RPC

Die Client-Server-Architektur ist bei netzbasierten Anwendungen das strukturelle Gliederungsprinzip. Dahinter steht die Idee, dass verschiedene Rechner unterschiedliche Dienste erbringen können, welche wiederum von anderen Rechnern und Benutzern eingesetzt werden. Um dieses Konzept auf einer Ebene zu realisieren, die nicht nur dem Benutzer der Anwendung, sondern auch deren Programmierer einen transparenten Einsatz von Serverdiensten ermöglicht, wurden *Remote-Procedure-Call-Systeme (RPC)* entwickelt. Sie erlauben die programmiersprachliche Benutzung von Objekten und Methoden über Rechnergrenzen hinweg. Dabei werden Probleme der Lokalisierung von Diensten, Parameterübertragung und Netzwerkzuverlässigkeit je nach eingesetztem System von diesem transparent gehandhabt. In der Folge kann der Anwendungsprogrammierer die angebotenen Dienste einsetzen wie jede Routine einer lokal vorhandenen Programmbibliothek. Realisiert wird dies durch sog. Proxies, also Platzhalter, die anstelle der funktionserbringenden Komponente vom Anwendungsprogramm aufgerufen werden. Diese Proxies übersetzen den Aufruf in ein internes Format und übertragen das daraus entstandene Datenpaket über eine Netzwerkverbindung an den Server. Auf Seiten des Servers wird dieser Prozess nun umgekehrt, d.h. der Aufruf wird dekodiert und an die Zielkomponente geleitet. Entsprechend erfolgt nach Ausführung der geforderten Funktionen die Rückübertragung von Ergebnissen. Beispiele für derartige Systeme sind CORBA, RMI oder DCOM sowie Elemente der .NET-Plattform (vgl. Haase 2001).

4.7.3 Web Services

Große Hoffnungen werden derzeit in die *Web Service*-Technologie (vgl. W3C 2002) gesetzt. Ähnlich den o.g. RPC-Systemen bieten sie eine Möglichkeit, Methodenaufrufe über ein Netzwerk zu tätigen. Die Stärken des Web Service-Konzepts liegen jedoch darin, dass die Aufruf- und Antwortdaten in einem standardisierten XML-Format übertragen werden (*SOAP*, vgl. W3C 2003), was den Einsatz über Grenzen von Programmiersprachen und auch Betriebssystemen hinweg fördert.

Zudem kann dadurch auf weitere Standards, wie XML-Encryption und XML-Signature zurückgegriffen werden, um Transaktionen zu sichern und die Benutzerrechte zu verifizieren. Für Web Services stehen mit *WSDL* (vgl. W3C 2001a) und *UDDI* (vgl. OASIS 2003) Beschreibungsformate zur Verfügung, die Information über syntaktische Eigenschaften und Protokollbindungen des Dienstes darstellen und diesen in Verzeichnissen listen können. Über diese Metainformationen kann ein Dienst von potenziellen Anwendern ermittelt und eingesetzt werden. Das von den Web Service-Initiatoren anvisierte Ziel ist, dass in Zukunft eine Vielzahl von Diensten im Internet angeboten werden, und Anwendungen diese dynamisch je nach aktuellen Anforderungen integrieren. Mit einer starken Verbreitung von Web Services ist in Zukunft auch deshalb zu rechnen, weil große Teile von Microsofts .NET-Technologie auf Web Services aufbauen und diese zudem eine Schnittstelle zwischen den beiden Integrationsplattformen .NET und J2EE darstellen (vgl. Abschnitt 4.7.4).

4.7.4 Komponentenintegration mit J2EE und .NET

Mit *J2EE* (vgl. Sun 2003) hat sich in den letzten Jahren eine Plattform etabliert, die konsequent auf die Mächtigkeit von Java baut. J2EE verbindet verschiedene Konzepte wie z.B. Web Services, Servlets und EJB (Enterprise Java Beans) zu einer Plattform, in deren Rahmen diese Komponenten zu komplexen Systemen integriert werden können. Die Kernidee dieses Integrationsansatzes besteht darin, dass die Produktentwicklung sich auf die tatsächlichen Kompetenzen eines Unternehmens beschränken kann, während Standardfunktionalität in Form von einzelnen Komponenten zugekauft und problemlos eingebunden werden kann (z.B. über EJB). Zusätzlich werden erweiterte Funktionalitäten wie Transaktionssicherheit für Komponenten systemseitig bereitgestellt und erleichtern somit die Entwicklung von Anwendungen.
Dazu ist in J2EE eine Application-Server Schnittstelle definiert (vgl. 4.7.5), die den einzelnen Komponenten den Zugriff auf Systemressourcen oder andere (auch auf anderen Application-Servern liegende) Komponenten ermöglicht. Da sich diese Schnittstelle mittlerweile zu einem Industriestandard etabliert hat, können die in diesem Rahmen eingesetzten Komponenten relativ einfach zwischen unterschiedlichen Systemen migriert werden.

Einen ähnlich umfassenden Anspruch erhebt Microsoft mit seiner neuen .NET-Technologie. .NET stellt die Weiterentwicklung der *Distributed Internet Applications Architecture (DNA)* dar und bietet zusätzlich zur bereits angeführten Programmiersprachen-Unabhängigkeit (4.7.1) auch entsprechende Funktionalität zur Netzwerkintegration. Im Vordergrund stehen hier Web Services und allgemein der Einsatz von XML an der Schnittstelle zwischen unterschiedlichen Komponenten. Eingebettet in diesen Kontext ist die Weiterentwicklung von ASP (vgl. 4.6.3), ASP.NET, das nun zudem nicht mehr skriptorientiert ist, sondern ebenfalls vorkompilierten Code zur Generierung von dynamischen Webseiten einsetzt, was deutliche Geschwindigkeitsvorteile gegenüber ASP bedeutet.

J2EE und .NET sind zwei Technologien, die zu einem großen Teil die gleichen Ziele haben. Beide bieten eine Integrationsplattform für eigenständige Komponenten wobei ebenfalls bei beiden die Netzwerkanbindung und –Transparenz ein wesentliches Merkmal darstellen. Während .NET aber eher den Einsatz verschiedener Programmiersprachen fokussiert, liegt das Gewicht bei J2EE eher auf der Plattformunabhängigkeit der Systeme. Folgerichtig sind .NET-Anwendungen heute im Wesentlichen auch nur auf Microsoft-Systemen einsetzbar, zeichnen sich dafür aber allgemein durch eine etwas bessere Performanz aus.

4.7.5 Application- und Integration Server

Ist mit J2EE auch eine *Application-Server* Schnittstelle spezifiziert, so kann dieser Begriff dennoch nicht auf die Java-Plattform alleine reduziert werden. Auch anderer Technologien (z.B. .NET) stellen die Basis für verschiedene Application-Server dar. Vielmehr bezeichnet dieser Begriff ein Konzept für die Entwicklung und den Einsatz von Business Anwendungen, bei dem verschiedene mehr oder weniger unabhängige Komponenten integriert werden und sich dem Benutzer als eine Einheit präsentieren. Dies kann von einzelnen Programmbausteinen reichen, bis hin zur Integration ganzer Applikationen zu einem komplexen Gesamtsystem (*Enterprise Application Integration*). In diesem Zusammenhang wird dann auch von *Integration Servern* gesprochen.

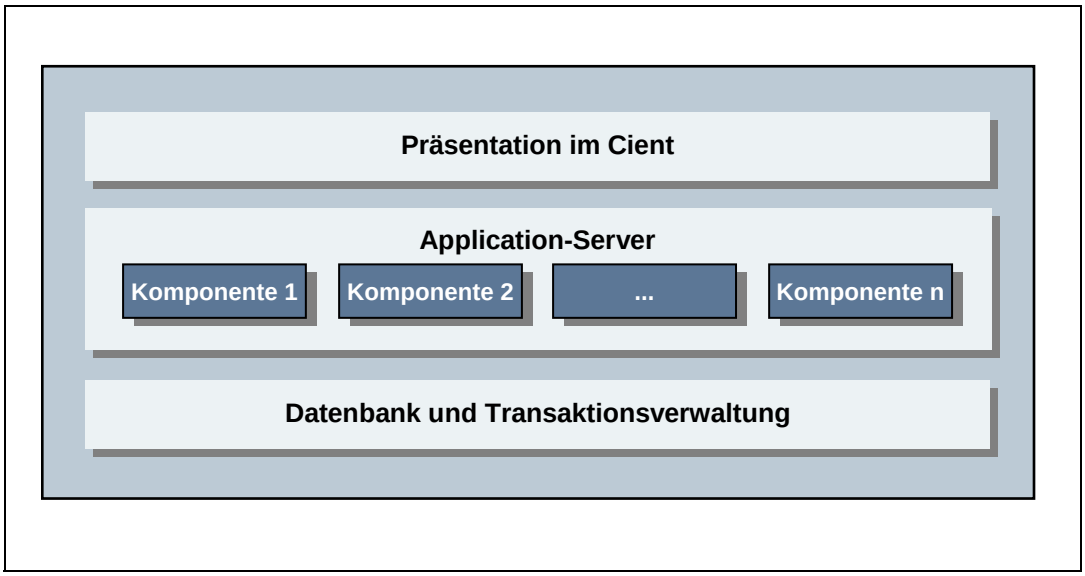

Abbildung 4-12: Einsatz von Application-Servern

Zentrale Eigenschaft von Application-Servern ist, dass sie den Aufbau der Anwendung in einer drei-Ebenen Struktur unterstützen, durch die die Bereiche Datenhaltung, Anwendungslogik und Präsentationslogik einfach voneinander getrennt werden können. Die Datenhaltung sowie Aufgaben der Transaktionsverwaltung wer-

den an Datenbanksysteme bzw. Transaktionsmonitore delegiert, wohingegen die Präsentation über Webbrowser oder spezielle Clients für den jeweiligen Application-Server erfolgt. Daher kann sich die Entwicklungsarbeit auf die tatsächliche Anwendungslogik beschränken, was wiederum den Aufwand für die Wartung der auf dem Application-Server betriebenen Anwendungen deutlich reduziert und geringe Hard- und Softwareanforderungen an den Client stellt.

4.7.6 Groupware

Um nicht nur verteilten Applikationen sondern auch verteilt arbeitenden Benutzern den Zugang zu gemeinsamen Ressourcen zu ermöglichen und ihre Kommunikation untereinander zu erleichtern, existieren verschiedene Softwaresysteme. Sie werden gemeinhin als *Groupware* bezeichnet.
In diese Kategorie fallen unter anderem einfache Systeme zur Verwaltung von Gruppenkalendern, ebenso aber auch Anwendungen, die räumlich verteilten Benutzern die Teilnahme an einer Konferenz ermöglichen. Dabei werden Funktionalitäten zur Verfügung gestellt, die den Austausch von Dokumenten ermöglichen oder ein gemeinsames Skizzieren von Sachverhalten auf einer virtuellen Tafel unterstützen. Für die direkte Kommunikation zwischen den Benutzern werden (teilweise multimediale) Foren eingesetzt, die je nach System synchrone oder asynchrone Kommunikation unterstützen, d.h. die gleichzeitige Anwesenheit der Teilnehmer im Forum erfordern bzw. neben der Freiheit des Raumes auch die freie Wahl des Nutzungszeitpunktes erlauben.
Derartige Funktionalität existiert bereits seit geraumer Zeit, jedoch hauptsächlich in Form von einzelnen, nicht miteinander gekoppelten Tools. Das ideale Groupware-System dagegen integriert unterschiedliche Einzelfunktionalitäten wie Gruppenkalender, Dokumentenarchive, Foren etc. und erweitert sie um zusätzliche Funktionalität. Dabei können Kommentare zu hinterlegten Materialien verwaltet oder Gruppenmitglieder bei Änderungen automatisch benachrichtigt werden. Als Beispiele für Systeme, die eine solche Integration bieten, seien hier Lotus Notes (http://www-306.ibm.com/software/lotus/) und das rein web-basierte BSCW (Basic Support for Cooperative Work, http://www.bscw.de) genannt. Der Vorteil der Integration besteht darin, dass verschiedenartige Informationen zu Projekten oder Arbeitspaketen zentral verfügbar sind und zudem eine lückenlose Protokollierung von Modifikationen möglich ist. Letzteres ermöglicht einen schnellen Überblick über Status und Fortschritt der Arbeit und erlaubt Entscheidungen und Entwicklungen zu einem späteren Zeitpunkt nachzuvollziehen.

4.8 Zusammenfassung

In den vorangegangen Abschnitten wurden verschiedene Technologien zur informationstechnischen Realisierung und Unterstützung des Information Workflow vorgestellt. Im Folgenden soll nun zusammenfassend aufgezeigt werden, wie diese bisher einzeln betrachteten Technologien zusammenhängen und wo sie sich ergänzen (vgl. Abbildung 4-13).

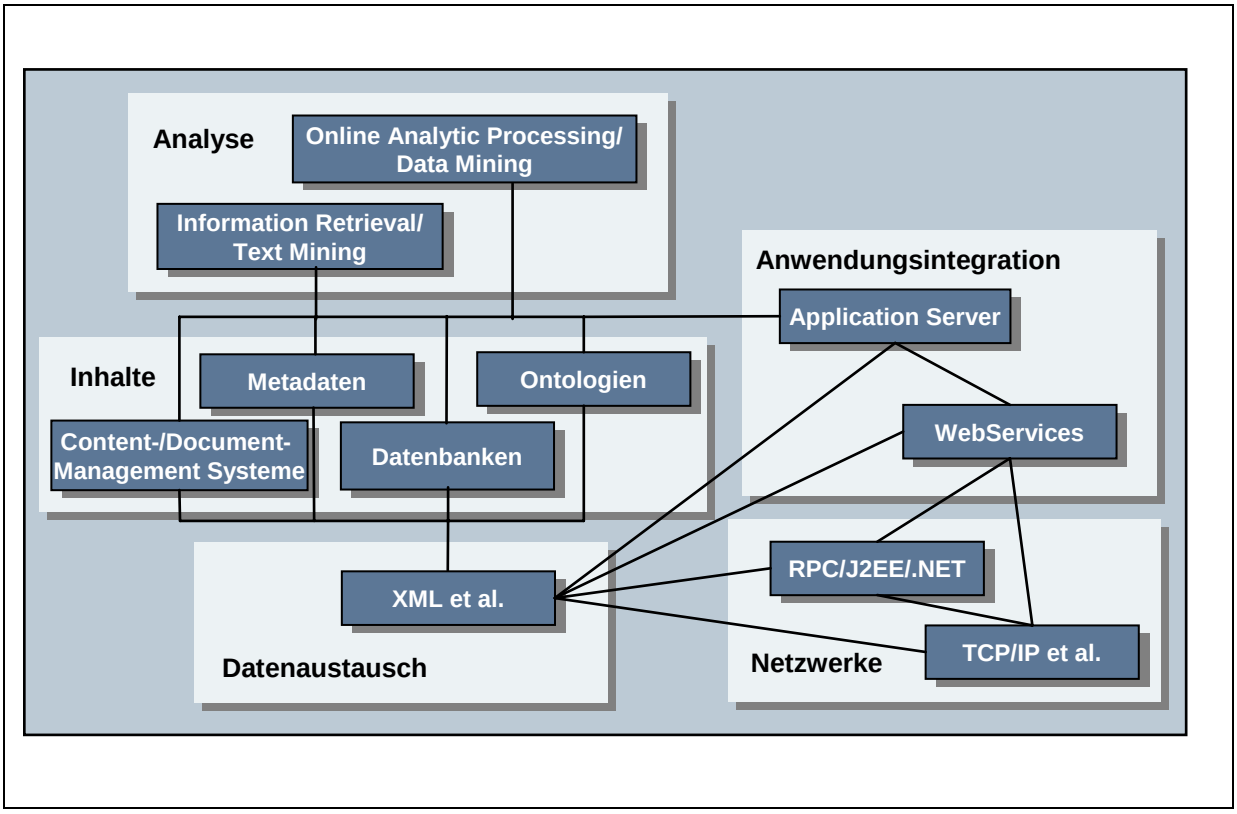

Abbildung 4-13: Zusammenwirken der einzelnen Technologien

Zentrale Bestandteile des Information Workflow sind die Informationsinhalte selbst. Sie werden im Zuge des Verarbeitungsprozesses erzeugt, bearbeitet und schließlich archiviert, ggf. in mehreren Iterationen. Der Zugriff auf Informationen und zugrunde liegende Daten erfolgt dabei meist über *Content*- oder *Document-Management-Systeme* bzw. im Rahmen von Anwendungsprogrammen direkt über *Datenbanken* (relational, objektrelational, hierarchisch, ...). Die Speicherung der Daten in solchen Systemen ist zunehmend von dauerhafter Natur. Dies liegt zum einen darin begründet, dass die Preise für Speicherkapazität kontinuierlich sinken und daher der Aufwand für das Filtern und Aussortieren nicht mehr benötigter Daten in einem immer schlechteren Verhältnis zum erhöhten Speicheraufwand steht. Des Weiteren bieten umfangreiche Datensammlungen offensichtlich eine bessere Grundlage für Analysemethoden wie *Online Analytic Processing* oder *Data/Text Mining* und stellen somit die Voraussetzung für ein tiefer gehendes Verständnis von nichttrivialen Zusammenhängen.

Die Schwierigkeit bei Speicherung und Zugriff auf umfangreichen Datenmengen besteht darin, aus dem gesamten Datenbestand die jeweils relevante Teilmenge zu identifizieren und einen effizienten Zugriff auf diese zu ermöglichen. Da verschiedene Anwendungen ihre Daten unterschiedlich strukturieren, ist es oft

schwer, eine übergreifende Katalogisierung und Suchfunktionalität anzubieten. Nach derzeitigem Stand der Technik ist aber eine vollautomatische Verarbeitung natürlichsprachlicher Texte mittels Hard- und Software und damit eine exakte Klassifizierung von Inhalten noch nicht möglich. Beim *Information Retrieval* bzw. *Text Mining* benutzt man daher Ansätze wie automatische Metadatenextraktion, Named Entitiy Recognition und darauf aufbauend *Ontologien* und Begriffsnetze, um diese fehlende exakte Methode zu kompensieren. Speicherung und Austausch der so ermittelten Metadaten erfolgt meist wieder in Form von XML (beispielsweise RDF) um eine einfache Zugriffsschnittstelle für weiterverarbeitende Anwendungskomponenten zu bieten.

Metadaten spielen aber auch im Bereich der Datenbanken eine entscheidende Rolle, also dort, wo die exakte Struktur der Daten bereits bekannt ist. Insbesondere bei der Integration von verteilten Datenbanksystemen (dazu zählt auch das sog. "Schema Matching") sind Informationen über die Struktur der Daten und beispielsweise Häufigkeiten von Werten oder Kardinalitäten von Attributen wichtig, um ein leistungsfähiges Datenbanksystem zu erhalten. Auch hier erfolgen Speicherung und Weitergabe solcher Metadaten aus o.g. Gründen meist in *XML*-basierten Formaten.

Der Begriff Information Workflow macht deutlich, dass Informationen bzw. die ihnen zugrunde liegenden Daten zwischen einzelnen Verarbeitungsstationen „fließen“ und damit von potenziell verschiedenen Personen in unterschiedlichen Anwendungen genutzt werden. Dass zunehmend Beschränkungen hinsichtlich Raum und Zeit der Datenverarbeitung fallen wird durch den vermehrten Einsatz von *Netzwerken* und verteilten Informationssystemen ermöglicht.

Um zusätzlich einheitliche Benutzeroberflächen sowie vereinfachte Administration zu ermöglichen, werden zunehmend unterschiedliche Anwendungen und Dienste mittels *Application-Server*-Technologien integriert. Diese wiederum bauen auf verteilten Systeme, da neben der Verarbeitung von Informationen durch Menschen auch die Leistungserbringung durch Rechner mehr und mehr in Netzwerken organisiert wird. Grundlage für die Integration verschiedener Anwendungen sind wieder Schnittstellen- und Datenbeschreibungen in standardisierten Formaten wie dem grundlegenden *XML*-Format, darauf aufbauend SOAP und UDDI und schließlich Highlevel-APIs wie *Web Services* oder das *.NET-Framework*.

5 Technologieübersicht für die Themenfelder im Projekt nova-net

Während auf den vorangegangenen Seiten ein allgemeiner Überblick über Technologien zur Be- und Verarbeitung von Informationen gegeben wurde, soll im Folgenden näher darauf eingegangen werden, welche dieser vorgestellten Technologien im Rahmen des Forschungsprojekts *nova-net* von besonderem Interesse sind. Die Darstellung orientiert sich dabei an den drei Schwerpunktthemen von *nova-net*, „Trendmonitoring im Szenario-Management“, „Life Cycle e-Valuation“ und „Lead User Integration“.

5.1 Trendmonitoring im Szenario-Management

5.1.1 Überblick über das Themenfeld

Unternehmen, die sich in heutigen, sich schnell verändernden, Märkten behaupten wollen, sind auf verlässliches Orientierungswissen angewiesen. Die Erschließung neuer Marktsegmente und die erfolgreiche Platzierung neuer Produkte in diesen Märkten ist eine immer schwieriger werdende Aufgabe. Die rasante Entwicklung neuer Technologien und die stetig steigenden Ansprüche von Kunden erschweren die Einschätzung von zukünftigen Produkterfolgen.

Die Untersuchung neuer Marktsegmente und die Planung von neuen Produkten muss aus diesen Gründen durch geeignete Informationsquellen und eine ausgereifte Methodik unterstützt werden, die einen Überblick über die zahlreichen Einflussfaktoren, die das Untersuchungsfeld bestimmen, bieten. Es muss ein Gesamtbild geschaffen werden, das den Untersuchungsgegenstand auf einen Blick präsentiert, jedoch ebenso in einzelne, nachvollziehbare Einflussfaktoren heruntergebrochen werden kann.

Diese Anforderungen erfüllt das Szenario-Management. Die Erfassung des Untersuchungsgegenstandes durch Szenarien und die Beobachtung von ausgearbeiteten Szenarien über längere Zeiträume hinweg (Szenario-Monitoring) sind Instrumente, die in den frühen Phasen des Innovationsmanagements (Orientierung, Ideenmanagement) ihren Einsatz finden. Die Einsatzmöglichkeiten von Szenarien sind vielfältig, da sie im Prinzip für die Beschreibung beliebiger Fragestellungen verwendet werden können. Die Kunst liegt in der Identifizierung und Auswahl von Einflussfaktoren, die das gegebene Themenfeld bestimmen.

Ziele des Szenario-Managements im Rahmen des Innovationsmanagements sind

- die Identifizierung von lukrativen und erfolgsversprechenden Marktsegmenten,
- die Einschätzung von Erfolgsaussichten für Produktideen

- und das Monitoring von Erfolgsaussichten über einen definierten Zeitraum hinweg.

Entscheidend für den erfolgreichen Einsatz ist dabei die Bündelung von unternehmensinternem sowie unternehmensexternem Wissen. In diesem Sinne ist das Management von Szenarien stark mit dem Themengebiet des Wissensmanagements verbunden. Einzelne Aspekte des Wissensmanagements, die in Bullinger 2001 anschaulich dargestellt werden und in diesem Zusammenhang beachtet werden müssen, sind folgende:

- Wissenserwerb
  Sowohl die Experten eines Kernteams, das Szenarien erarbeitet, als auch unternehmensexterne Experten steuern wertvolles Wissen für das Szenario-Management bei. Der Erwerb dieses Wissens ist mitnichten trivial. Szenario-Management-Workshops müssen professionell moderiert werden. Auch die Gestaltung von Fragebögen, die bei der Befragung unternehmensexterner Experten zum Einsatz kommen, müssen sorgfältig vorbereitet werden.
- Wissensspeicherung
  Das extrahierte Wissen muss in geeigneter Form gespeichert werden. Im Rahmen des Szenario-Managements werden zum Beispiel Einflussfaktoren aufgestellt, die eine gegebene Fragestellung bestimmen.
- Wissensaufbereitung
  Das in geeigneter Form gespeicherte Wissen sollte anschließend verarbeitet und für die Weiterverwendung aufbereitet werden. Szenarien, Einflussfaktoren und ihre Ausprägungen müssen so präsentiert werden, dass sie das gegebene Themengebiet auf einem Blick darstellen, jedoch die Verfolgung einzelner Informationsstränge nicht ausschließen.
- Wissensnutzung
  Das erarbeitete Wissen soll im Unternehmen dazu verwendet werden, um strategische Entscheidungen auf Basis solider Informationen treffen zu können. Das aufbereitete Wissen muss Entscheidungsträgern unter Berücksichtigung ihrer Kenntnisse und Fähigkeiten präsentiert werden.
- Wissensaktualisierung
  Erarbeitete Szenarien stellen zwar eine wichtige Entscheidungshilfe für das Management eines Unternehmens dar, sie bieten jedoch nur eine Einschätzung die ihre Gültigkeit schnell verlieren kann. Turbulente Märkte und neue Technologien sorgen dafür, dass Fragestellungen in immer kürzer werdenden Abständen neu beantwortet werden müssen. Um diesem Problem entgegenzutreten, muss das gesammelte Wissen aktualisiert, und damit ein Monitoring mit Szenarien ermöglicht werden.

Wie aus o.g. Punkten ersichtlich ist, nehmen Erstellung und Management von Szenarien vielfältige Ressourcen eines Unternehmens in Anspruch. Zum einen müssen diverse Informationsarten und Informationsquellen, wie Technologieentwicklungen, Konsumverhalten, Produktionsdaten, usw. in Betracht gezogen wer-

den, zum anderen ist ein erheblicher personeller Aufwand für die Szenarienbildung notwendig. Diese Ressourcen sind in großen internationalen Unternehmen räumlich und zeitlich verteilt. Aus diesem Grund ist für Entwicklung und Monitoring von Szenarien eine informationstechnische Unterstützung notwendig, die sowohl den personellen Aufwand reduziert, als auch die effiziente Einbindung von verschiedenen Informationsquellen realisiert.

5.1.2 Ziele der informationstechnischen Unterstützung im Themenfeld

Die Entwicklung von Szenarien im Rahmen eines Innovationsprojektes im Unternehmen wird optimalerweise von einem kompetenten Kernteam durchgeführt. Das Kernteam besteht in der Regel aus Entscheidungsträgern verschiedener Bereiche des Unternehmens, wie z.B. Entwicklung, Marketing oder Produktion. Die Teilnehmer des Kernteams müssen mit unternehmensinternen und unternehmensexternen Informationen versorgt werden, wobei die Präsentation der Daten den Fähigkeiten und Kenntnissen der einzelnen Teammitglieder angepasst werden muss.

Zusätzlich müssen oft unternehmensexterne Experten in die Entwicklung von Szenarien eingebunden werden. Ihre Meinung zu einzelnen Einflussfaktoren, die das Themenfeld bestimmen, kann zum Beispiel per Fragebogen im Rahmen einer Delphi-Befragung erhoben werden.

Das Wissen des Kernteams und der unternehmensexternen Experten soll schließlich dazu verwendet werden, die richtigen Einflussfaktoren für die Szenarioentwicklung festzulegen, wobei die Bildung von Szenarien in der Regel vom Kernteam übernommen wird. Durch die Abschätzung der zukünftigen Entwicklung von Einflussfaktoren ergeben sich anschließend verschiedene Zukunftsszenarien. Der Auswahlprozess für Einflussfaktoren, die Bestimmung zukünftiger Entwicklungsrichtungen und die Darstellung von Szenarien sind wichtige Aspekte im Rahmen des Szenario-Entwicklungsprozesses, die durch die geeignete informations- und softwaretechnische Unterstützung erheblich vereinfacht werden können.

Ziele der informationstechnischen Unterstützung sind im Einzelnen:

1. Unterstützung des Kernteams

   - Die strukturierte Entwicklung von Szenarien soll idealerweise durch eine umfassende Softwarelösung unterstützt werden, die das Kernteam, basierend auf einer ausgereiften Methodik, durch den Entwicklungsprozess leitet. Diese Softwarelösung soll alle Aspekte der Szenarienentwicklung abdecken.
   - Mitglieder des Kernteams sollten jederzeit auf benötigte Informationen aus ihrem Kompetenzbereich zugreifen können. Dementsprechend soll zum Beispiel der Produktionsleiter eines Unternehmens auf Produktionskennzahlen zurückgreifen können, um seine Einschätzung zu untermauern. Er-

forderlich ist in diesem Zusammenhang der Zugriff auf unternehmensinterne Datenquellen wie Dokumentenarchive und Datenbanken.

- Das Kernteam sollte durch unternehmensexterne Informationsquellen unterstützt werden. Hierbei sind in erster Linie Internetressourcen, wie Online-Medien, Foren und Newsgroups gemeint, die entsprechend aufbereitet entscheidungsunterstützend dienen können.

2. Einbinden unternehmensexterner Experten

   Das Wissen unternehmensexterner Experten ist in vielen Fällen erforderlich, um die zukünftige Entwicklung von Einflussfaktoren einzuschätzen. Bei der Einbindung externen Expertenwissens in die Szenarienentwicklung muss vor allem die räumliche und zeitliche Verteilung von Experten berücksichtigt werden. Aus diesem Grund müssen bewährte Befragungstechniken mit den entsprechenden Internettechnologien verknüpft eingesetzt werden. Unterstützt durch E-Mail und Web-Fragebogen kann zum Beispiel eine Delphi-Befragung erheblich schneller durchgeführt und ausgewertet werden.

5.1.3 Technologien und Werkzeuge zur informationstechnischen Unterstützung

Das Management von Szenarien sollte, wie bereits erwähnt, durch ein System unterstützt werden, das alle Aspekte der Szenarienentwicklung abdeckt. Zusätzlich sollte es möglich sein, den Entwicklungsprozess an den speziellen Anforderungen des Unternehmens auszurichten oder sogar neue Methoden und Verfahren in das System zu integrieren.

Es existieren bereits einige Softwarelösungen auf dem Markt, die verschiedene Methoden der Szenarienentwicklung umsetzen. Der Großteil dieser Werkzeuge verwendet die so genannte Szenario-Technik. Im Rahmen der Szenario-Technik wird ein Themenfeld durch eine Menge von Einflussfaktoren beschrieben. Aus den Eintrittswahrscheinlichkeiten verschiedener Ausprägungen eines Einflussfaktors oder aus der gegenseitigen Beeinflussung von Einflussfaktor-Ausprägungen können auf verschiedene Art und Weise Szenarien abgeleitet werden. Die Konsistenzanalyse betrachtet die Verträglichkeit von Einflussfaktor-Ausprägungen, während die Cross-Impact-Analyse die Eintrittswahrscheinlichkeiten von Einflussfaktor-Ausprägungen und ihre Wechselwirkungen analysiert (Gausemeier 1996).

5.1.3.1 Szenario-Management Werkzeuge

Im Folgenden sollen einige dieser Werkzeuge und ihre Eigenschaften vorgestellt werden.

CIM 8.0

Das Werkzeug CIM 8.0 (vgl. CIM 2003) unterstützt die Durchführung der Cross-Impact-Analyse im Rahmen von Workshops. Die Cross-Impact-Analyse wird in drei Schritten durchgeführt.

- Im ersten Schritt werden die für das Themenfeld relevanten Einflussbereiche, Einflussfaktoren und mögliche Einflussfaktor-Ausprägungen von einem personell geeignet zusammengesetzten Team festgelegt.
- Im zweiten Schritt wird die gegenseitige Beeinflussung der einzelnen Einflussfaktoren-Ausprägungen bestimmt. Als Ergebnis entsteht eine Cross-Impact-Matrix, die für jeden Einflussfaktor angibt, mit welcher Stärke und Richtung er jeden anderen Einflussfaktor beeinflusst.
- Im letzten Schritt werden aus den Einflussfaktoren und aus ihren Beziehungen wahrscheinliche Szenarien berechnet. Die Auswertung der Szenarien wird durch den Einsatz verschiedener Analysegraphiken ermöglicht.

Dieses Werkzeug baut auf der Szenario-Technik auf, bietet jedoch nur die Cross-Impact-Analyse an, die nur eine einzelne Methode der Szenario-Technik darstellt. Ein gravierender Nachteil der Cross-Impact-Analyse ist der enorme Aufwand, der für die Entwicklung der Cross-Impact-Matrix benötigt wird. Die Zahl der abzugebenden Schätzungen für die Matrix steigt quadratisch mit der Zahl der Einflussfaktoren.

INKA 3

Das Werkzeug INKA 3 (vgl. INKA) stützt sich ebenfalls auf die Szenario-Technik und unterstützt die so genannte Konsistenz-Analyse. Diese wird in drei Schritten durchgeführt.

- Im ersten Schritt werden Einflussfaktoren und ihre möglichen Ausprägungen für das Themenfeld bestimmt. Die Erarbeitung und Speicherung der Daten erfolgt optimalerweise ebenfalls durch eine Gruppe von Experten.
- Im zweiten Schritt werden Konsistenzabschätzungen abgegeben, die festlegen, ob Einflussbereich-Ausprägungen zusammen auftreten können. Diese Konsistenzen können in einer Konsistenz-Matrix gespeichert werden.
- Schließlich können aus den Konsistenzbeziehungen konsistente Szenarien abgeleitet werden. Die Szenarien werden algorithmisch durch Ausschluss inkonsistenter Einflussfaktor-Kombinationen berechnet.

Die Konsistenz-Analyse ist mit deutlich weniger Aufwand als die Cross-Impact-Analyse durchführbar. Einerseits muss für zwei Einflussbereich-Ausprägungen nur eine „Ja-Nein“ Entscheidung getroffen werden, andererseits muss nur die halbe

Matrix ausgefüllt werden. Die Konsistenz-Analyse ist jedoch nur eine einzelne Methode im Rahmen der Szenario-Technik.

Szeno-Plan

Szeno-Plan (Szeno 2004) baut ebenfalls auf die Szenario-Technik auf und unterstützt sowohl die Konsistenz-Analyse als auch die Cross-Impact-Analyse. Nach Ermittlung der Einflussfaktoren und ihren Ausprägungen kann der Anwender beide Analysen durchführen. Die Softwarelösung bietet auch Schnittstellen zu gängigen Office Anwendungen, um zum Beispiel die erhaltenen Ergebnisse in einer Tabellen-Kalkulations-Anwendung weiterverarbeiten zu können.

Szenario-Manager

Der Szenario-Manager (SzenMan 2004) ist ein Werkzeug, das die größte Funktionalität in der Reihe der Szenario-Technik-Anwendungen bietet. Die Szenarien basieren ebenfalls auf eine Menge von Einflussfaktoren sowie deren Ausprägungen. Auf Grund dieser Daten werden neben der Konsistenz- und Cross-Impact-Analyse auch weitere Verfahren, wie die Clustering oder die Multidimensionale Skalierung, angeboten.

Im Gegensatz zu den bereits vorgestellten Tools verwendet der Szenario-Manager eine Client-Server-Architektur. Die Clients dürfen lediglich online von den Anwendern bedient werden. Berechnungen finden auf dem Server der Betreiber statt. Diese Art der Verwendung von Clients und Servern stört im Bereich des Innovationsmanagements sicherlich die Akzeptanz der Software. Sicherheitsbewusste Kunden werden ihre Entwicklungs- und Innovationsdaten nur sehr ungern in fremde Hände geben und auf fremden Servern speichern.

5.1.3.2 Anforderungen für eine umfassende Szenario-Management-Lösung

Die vorgestellten Werkzeuge unterstützen einige Aspekte des Szenario-Managements, indem sie verschiedene Methoden im Rahmen der Szenario-Technik umsetzen. Das Management von Szenarien wird von der Szenario-Technik jedoch keineswegs vollständig erfasst. Ein System, das das Management von Szenarien umfassend unterstützen will, sollte nicht nur eine Methodik bieten, sondern auch die Umsetzung der Methodik zur Entwicklung von Szenarien vollständig unterstützen. Die Szenario-Technik setzt voraus, dass Expertenwissen für die Erarbeitung der Einflussfaktoren vorhanden ist und jederzeit abgerufen werden kann. Das System muss jedoch unter Anderem verschiedene Informationsquellen bereitstellen, um die Experten bei der Erstellung von bestmöglichen Einschätzungen zu unterstützen. Hierzu gehören zum Beispiel verschiedene unternehmensinterne Informationsquellen, wie Produktionsdatenbanken, Informationen über Produktionsabläufe oder Kundendaten. Unternehmensexterne Informationsquellen, wie zum Beispiel Ressourcen aus dem World Wide Web, sollten ebenfalls zur Ver-

fügung gestellt werden. In Falle der Web-Ressourcen muss eine entsprechende Vorverarbeitung (Suche, Filterung, Ranking) stattfinden.

Ein weiterer Aspekt, der von keinem der vorgestellten Systeme unterstützt wird, ist die dauerhafte Verbindung von Informationsquellen mit Szenarien, um somit das Monitoring mittels Szenarien zu ermöglichen. Im Rahmen der Szenario-Technik wird ein Themenfeld durch Einflussfaktoren beschrieben. Mögliche Ausprägungen der Einflussfaktoren ergeben schließlich verschiedene künftige Szenarien. Welche Ausprägung für einen Einflussfaktor mit welcher Wahrscheinlichkeit eintritt, wird von Experten vorhergesagt. Einflussfaktoren, die maßgeblich von quantitativen Größen abhängen (z.B. Produktionskennzahlen), können jedoch automatisch durch die Aktualisierung der Ausgangsgrößen auf dem neuesten Stand gehalten werden. Hierzu sind natürlich geeignete Formeln und Schwellenwerte notwendig, um die Wahrscheinlichkeit der Einflussfaktor-Ausprägungen aus den Ausgangsgrößen zu berechnen. Für Einflussbereiche, für die Ausprägungswahrscheinlichkeiten auf diesem Wege aktuell gehalten werden, müssen nur einmal Experten konsultiert werden. Ihre Aufgabe besteht dann nicht in der wiederholten Bestimmung der Ausprägungswahrscheinlichkeiten, sondern in der Festlegung der Beziehung zwischen Ausgangsgrößen und Ausprägungswahrscheinlichkeiten. Sind die Formeln und Schwellenwerte, die diese Beziehung herstellen einmal erarbeitet, kann der entsprechende Einflussfaktor ohne weitere Beteiligung der Experten auf dem neuesten Stand gehalten werden.

Einer der größten Nachteile, den alle vorgestellten Werkzeuge aufweisen, ist die mangelnde Internetunterstützung. Das Internet und insbesondere das World Wide Web bieten zahlreiche Möglichkeiten für die effiziente Entwicklung von Szenarien. Zum einen kann das WWW als Informationsquelle dienen und somit verlässlichere Expertenaussagen ermöglichen. Zum anderen können das Internet und das WWW als Zugangs- und Übertragungsmedium dienen, um räumlich und zeitlich verteilte Anwender eines Szenario-Management Systems miteinander zu verbinden.

Ein letzter, jedoch nicht minder wichtiger Punkt ist die Architektur, die einem Szenario-Management-System zu Grunde liegen sollte. Der Großteil der verfügbaren Szenario-Management-Werkzeuge sind Einzelanwendungen, die weder sonderlich erweiterbar noch miteinander kombinierbar sind. Obwohl einige dieser Werkzeuge Import- und Exportschnittstellen mitbringen, wäre der Austausch von Zwischenergebnissen, wenn überhaupt möglich, mit zusätzlichem Transformationsaufwand und mit Informationsverlust verbunden. Aus diesem Grund sollte ein Szenario-Management-System aus sorgfältig konzipierten Modulen bestehen. Diese sollten das Hinzufügen weiterer Module und die Weitergabe der Zwischenergebnisse in definierten Formaten ermöglichen. Zusätzlich sollte die Systemarchitektur das verteilte Arbeiten über das Internet ermöglichen.

5.1.3.3 Wichtige Technologien für das Szenario Management

Im vorhergehenden Abschnitt wurden einige Anforderungen definiert, die von heutigen Szenario-Management-Werkzeugen nicht erfüllt werden. Für die Konzeption einer umfassenden Szenario-Management-Lösung unter Beachtung der aufgestellten Anforderungen sind verschiedene Technologien notwendig, die in diesem Abschnitt vorgestellt werden. Für jede Technologie wird erläutert, in welchem Zusammenhang sie für das Management von Szenarien eingesetzt werden kann.

Application Server Technologie

Wie bereits angedeutet, können nicht alle Anwender eines Szenario-Management-Systems vor einem Bildschirm versammelt werden, um dort gemeinsam das System zu verwenden. Obwohl Szenarien optimalerweise im Rahmen von Workshops von einem kleinen Kernteam entwickelt werden, sollten Mitglieder dieses Teams auch die Möglichkeit haben, das System einzeln und parallel zu verwenden. Dies ist für die Vorbereitung der Workshop-Treffen ebenso wichtig wie für die Betrachtung und Auswertung der fertigen Szenarien im Nachhinein.

Für die Umsetzung bietet sich deswegen eine Client-Server-Architektur an. Beliebig viele Clients sollten über das Unternehmensintranet oder über das Internet miteinander kommunizieren, und somit eine räumlich und zeitlich verteilte Nutzung des zugrunde liegenden Systems ermöglichen. Für die Umsetzung der Server-Komponente bietet sich in diesem Kontext die Application-Server-Technologie an.

Ein Application Server ist eine Softwarekomponente auf einem Computer, der eine bestimmte Funktionalität (Anwendung) anderen Computern (Clients) zur Verfügung stellt. Die Clients verwenden die angebotene Funktionalität des Servers, und verarbeiten die übermittelten Daten weiter (vgl. Abschnitt 4.7.5).

Webserver Technologie

Anwendern des Szenario-Management Systems sollte ein möglichst unkomplizierter Zugang zum System geboten werden. Dies kann erreicht werden, indem man zum Beispiel Web-Browser als Client-Software einsetzt. Dazu ist es notwendig, dass neben dem Application Server zusätzlich ein Web Server eingesetzt wird, und dass die Bedienoberfläche des Systems Web-Technologien verwendet, damit Inhalte in Web-Browsern angezeigt werden können. Zu diesen Web-Technologien gehören HTML, die Cascading Stylesheets, JavaScript aber auch Java Applets oder Macromedia Flash.

In diesem Fall wird ein Großteil des Systems auf der Serverseite implementiert. Die Aufgabe des Clients besteht lediglich in der Interaktion mit dem Benutzer.

Content Management Technologie

Application Server, Web-Server und Web-Browser formen eine flexible Infrastruktur für ein Szenario-Management System. Es gibt jedoch weitere Technologien, die sich weniger für die Bildung einer Systeminfrastruktur eignen, sondern der Verarbeitung von Inhalten dienen. Content-Management-Systeme erleichtern die Erstellung und Verwaltung von Inhalten, indem sie den Erstellungs- und Verwaltungsprozess in verschiedene Arbeitsschritte unterteilen. Bezogen auf die Inhalte bilden diese den so genannten Content-Lifecycle (vgl. Abschnitt 4.6.1). Kombiniert man dies mit der Client-Server-Technologie, dann können mehrere Clients gleichzeitig verschiedene Arbeitsschritte ausführen, und somit für eine effiziente Verarbeitung der Gesamtaufgabe sorgen. Im Rahmen des Szenario-Managements wäre zum Beispiel vorstellbar, dass eine Person oder Gruppe sich nur um die Sammlung von Einflussfaktoren für ein bestimmtes Themengebiet kümmert, während eine andere Gruppe sich der Erstellung von Szenarien widmet.

5.1.4 Szenario-Management Framework

Die vorangehenden Abschnitte haben verdeutlicht, dass momentan keine Lösung existiert, die das Szenario-Management umfassend unterstützt. Existierende Werkzeuge ermöglichen die Durchführung einer oder mehrerer Methoden der Szenario-Technik, vernachlässigen jedoch die informationstechnische Unterstützung für den gesamten Szenario-Erstellungsprozess.

Um diese Lücke zu füllen wird in diesem Abschnitt der Szenario Management Framework vorgestellt. Dieses Framework (vgl. Abbildung 5-1) präsentiert einen technischen Rahmen, indem Szenario-Management Anwendungen entwickelt werden können.

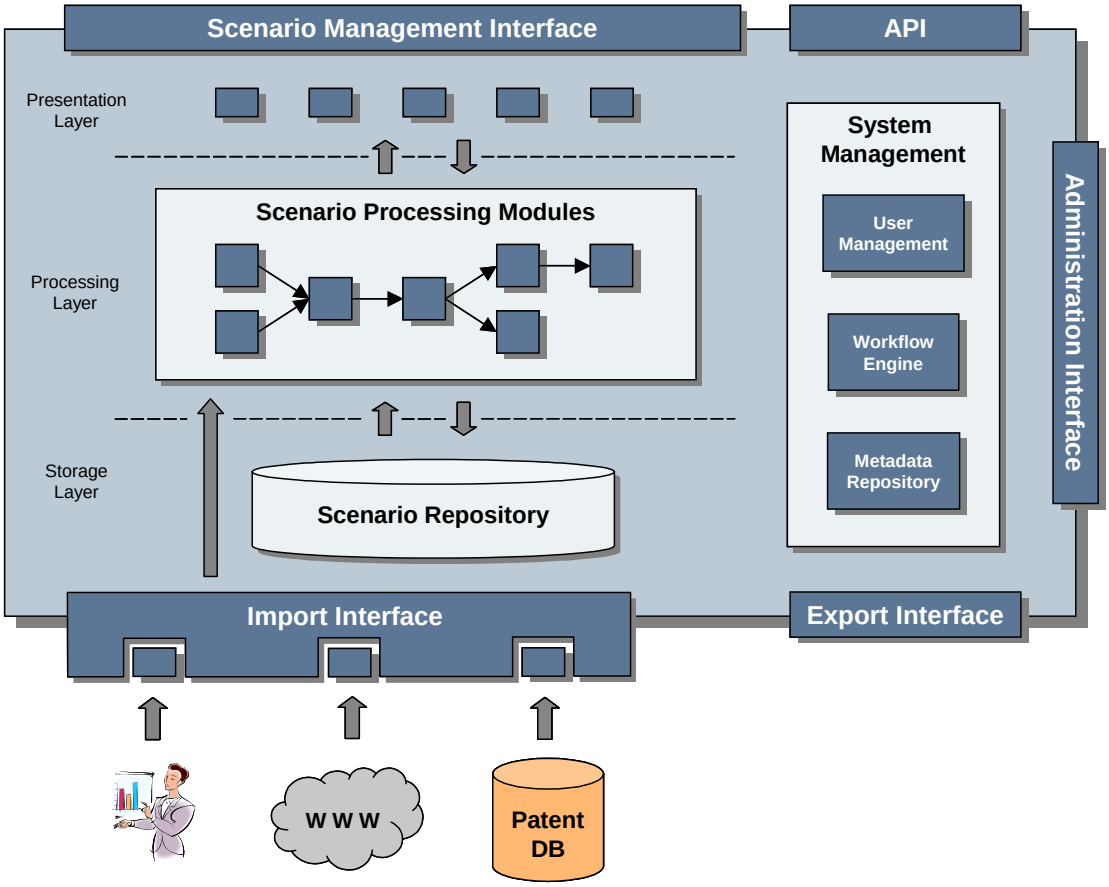

Abbildung 5-1: Szenario Management Framework

Die Hauptkomponenten des Frameworks sind in drei Schichten angeordnet. In der Speicherungsschicht sorgt das Scenario Repository für das Ablegen von Daten und Objekten (z.B. Einflussfaktoren, Szenarien, …), die im Szenario-Erstellungsprozess benötigt werden. Die Verarbeitungsschicht bietet Module (z.B. Einflussfaktor-Modul, Konsistenz-Analyse-Modul, …), die Daten der Speicherungsschicht verwenden, um eine breite Palette von Szenario-Management Aufgaben zu unterstützen. Aus den einzelnen Modulen können benutzerdefinierte Workflows gebildet werden, die eine flexible und effiziente Verarbeitung ermöglichen. Neue Module können dem System jederzeit hinzugefügt und mit anderen Modulen kombiniert werden. Die Präsentationsschicht stellt Objekte des Systems für die Benutzer dar.

Der Framework verfügt über verschiedene Schnittstellen. Die Import Schnittstelle ermöglicht das Einbinden verschiedener Informationsquellen (Experten, Daten-

banken, WWW, …) und unterstützt damit die Erstellung von Szenarien, die einerseits auf einer soliden Informationsbasis ruhen, andererseits laufend aktualisiert werden können. Die Szenario-Management Schnittstelle bietet eine graphische Benutzeroberfläche für Szenario-Manager, während die Administrationsschnittstelle den Zugang zu verschiedenen Systemmanagementfunktionen darstellt. Schließlich können Szenario-Management Daten mittels der Export-Schnittstelle und dem API anderen Systemen zur Verfügung gestellt werden.

Das Szenario Management Framework ist ein Modell, nach dem ein System für das Management von Szenarien entwickelt werden sollte. Es unterstützt alle Anforderungen, die in Abschnitt 5.1.3.2 aufgestellt wurden. Dementsprechend müssen bei der technischen Umsetzung des Frameworks die in Abschnitt 5.1.3.3 vorgestellten Technologien verwendet werden.

5.2 Life Cycle e-Valuation

5.2.1 Überblick über das Themenfeld

Produkte, Dienstleistungen und Produktsysteme werden während ihrer Entwicklung einer Vielzahl von Bewertungen unterzogen. Im Rahmen der Life Cycle e-Valuation steht die Bewertung von Nachhaltigkeitsaspekten im Mittelpunkt. Für Teilaspekte einer Nachhaltigkeitsbewertung sind Methoden der Umweltwirkungsbewertung verfügbar. Hierzu gehören beispielsweise das Life Cycle Management und das Life Cycle Assessment sowie Methoden der erweiterten Wirtschaftlichkeitsrechnung (siehe Abschnitt 2.2 sowie Lang et al. 2004).

Für viele der genannten Ansätze ist einerseits eine ganzheitliche Betrachtung des gesamten Lebenszyklus eines Produktes kennzeichnend. Andererseits fließen in die jeweilige Bewertungsmethodik umfangreiche exakte Daten zu einem Produkt oder einer Dienstleistung ein. Die Durchführbarkeit der Nachhaltigkeitsbewertung mit Hilfe dieser Methoden hängt somit direkt von der Verfügbarkeit exakter Daten zu einem Produkt oder einer Dienstleistung ab. In der Regel sind die Daten in der Phase des Projektmanagements (siehe Abbildung 2-1) in hinreichendem Umfang und Detaillierungsgrad vorhanden. Hier kann somit die Umweltwirkungsbewertung mit den oben angegebenen Methoden unmittelbar ansetzen. Anders sieht es in den frühen Phasen eines Innovationsprozesses aus. Insbesondere beim Ideenmanagement (Ideengewinnung, Ideenbewertung und Ideenauswahl), liegen solche Daten in der Regel nicht bzw. nicht in ausreichendem Detaillierungsgrad vor. Für diese Phasen müssen somit andere Ansätze zur Nachhaltigkeitsbewertung gefunden werden. Diese Ansätze müssen durch sehr geringe Anforderungen, was den Umfang und die Exaktheit der zugrunde liegenden Daten angeht, gekennzeichnet sein. Der frühzeitigen Nachhaltigkeitsbewertung in Innovationsprozessen kommt eine sehr große Bedeutung zu, da hiermit einzelne Ideen und Vorschläge identifiziert werden können, die in die Weiterentwicklung der Produktidee im Innovationsprozesses mit einfließen. Diese können dann entweder bei der Ausarbeitung der

Ideen berücksichtigt werden oder es kann im Extremfall vom kostenintensiven Weiterverfolgen solcher Ideen abgesehen werden.

Da in den frühen Phasen des Innovationsprozesses einem Unternehmen noch nicht ausreichend detaillierte, interne Daten für eine Nachhaltigkeitsbewertung mit Methoden der Umweltwirkungsbewertung zur Verfügung stehen, ist es nahe liegend entsprechende zusätzliche Informationen im Internet zu suchen. Ein Schwerpunkt der informationstechnischen Unterstützung für den Bereich Life Cycle e-Valuation liegt darum auf Suchstrategien, die es erlauben für die Nachhaltigkeitsbewertung zu einzelnen Produktideen relevante Information im Internet zu identifizieren und geeignet zu präsentieren.

Zusammenfassend kann der Bereich Life Cycle e-Valuation im Projekt *nova-net* also als die informationstechnisch unterstützte Nachhaltigkeitsbewertung von Produkten, Dienstleistungen und Produktsystemen betrachtet werden. Die Ziele der informationstechnischen Unterstützung werden im folgenden Abschnitt genauer erläutert.

5.2.2 Ziele der informationstechnischen Unterstützung im Themenfeld

Die anzustrebende informationstechnische Unterstützung für die frühen Phasen von Innovationsprozessen kann wie folgt charakterisiert werden: Auf der Grundlage weniger Basisdaten zu einem Produkt oder einer Dienstleistung werden im Internet Informationen gesucht, die eine einfache Nachhaltigkeitsbewertung ermöglichen. Einzelne Aspekte dieser Zieldefinition werden im Folgenden genauer erläutert.

Das Internet als breite und im Allgemeinen kostengünstig verfügbare Informationsbasis ist hier unabdingbar. Abhängig von den Ergebnissen von Umfeld- und Trendanalysen werden in Innovationsprozessen häufig Produktideen verfolgt, die wenig Zusammenhang zu der aktuellen Produktpalette eines Unternehmens aufweisen. Des Weiteren müssen aktuelle Entwicklungen und Diskussionen über die Umweltwirkungen von Produkten und deren Bezug zur konkreten Produktidee verfolgt werden. In solchen Fällen ist davon auszugehen, dass in unternehmensinternen Datenbanken sowie im Intranet eines Unternehmens kaum Informationen vorliegen, die einen nennenswerten Beitrag zur Nachhaltigkeitsbewertung des neunen Produktes bzw. der Dienstleistung leisten können. Als Konsequenz daraus muss eine wesentlich breitere Informationsbasis gewählt werden.

Beim Gegenstand der Suche kann es sich um einfache Dokumente, um Methoden aber auch um Experten handeln. Es sollte also möglich sein, Dokumente zu identifizieren, die etwas über Chancen und Risiken im Zusammenhang mit der Umweltwirkung eines Produkts, einer Dienstleistung, eines Werkstoffs, etc. aussagen und/oder Umweltgesichtspunkte im Zusammenhang mit einer Branche oder einem Produktfeld betrachten. Der Begriff *Dokument* umfasst hier sowohl einfache Web-

Seiten als auch Dokumente wie Zeitschriftenartikel und wissenschaftliche Publikationen, die häufig im PDF-Format im Internet veröffentlicht werden.

Ein weiteres Ergebnis einer Recherche im Internet können Verweise auf Methoden zur Nachhaltigkeitsbewertung sein. Auf diese Weise können Methoden identifiziert werden, die speziell auf den zu betrachtenden Bereich (neues Produkt oder Dienstleistung) zugeschnitten sind. Dies umfasst Methoden, die in Form von Softwarewerkzeugen verfügbar sind, Methoden, die in Form von (Web-) Services genutzt werden können, aber auch Methoden für die eine informationstechnische Unterstützung nicht vorgesehen und evtl. auch gar nicht möglich ist. Die Bewertung gefundener Methoden für die im konkreten Projekt angestrebte Nachhaltigkeitsbewertung muss ergänzend zur Suche erfolgen.

Darüber hinaus stellen Verweise auf Experten eine wichtige Ergebniskategorie für die Suche im Internet dar. In diesem Fall liefert die Suche selbst keinerlei Information, die direkt für eine Nachhaltigkeitsbewertung genutzt werden kann. Diese kann erst durch den Kontakt zu den identifizierten Experten verfügbar gemacht werden. Die Suche nach Experten für spezifische Themenfelder ist ebenso eine wichtige Aufgabe im Rahmen der Lead-User-Identifikation und –Integration. Sie wird darum erst im Abschnitt 5.3 genauer betrachtet.

Unabhängig vom genauen Ziel der Suche sollte sie mit möglichst wenig Expertenwissen durchgeführt werden können. Diejenigen, die im Rahmen des Innovationsprozesses eine Nachhaltigkeitsbewertung in frühen Phasen durchführen wollen sind in der Regel weder Nachhaltigkeitsexperten (die Akteure kommen vor allem aus den Bereichen Forschung/Entwicklung und Marketing) noch Experten für den Innovationsgegenstand. Letzteres gilt insbesondere bei Innovationen, die aus Unternehmenssicht eine Radikalinnovation darstellen. Das Vorhandensein umfangreichen Expertenwissens sollte demnach keine Voraussetzung für die erfolgreiche und zielgerichtete Suche nach Nachhaltigkeitsinformationen im Internet sein. Nichtsdestotrotz ist es denkbar, dass Expertenwissen unabhängig von den einzelnen Nutzern bereitgestellt wird und für die Nutzer transparent in den Suchprozess einfließt.

5.2.3 Technologien und Werkzeuge zur informationstechnischen Unterstützung

In diesem Abschnitt wird detailliert erläutert welche Technologien für die beschriebenen Suchaufgaben im Bereich Life Cycle e-Valuation genutzt werden können, welche Ziele jeweils erreicht werden und wo die Schwachstellen des jeweiligen Ansatzes liegen. Ergänzt wird die Erläuterung der Technologien jeweils um eine Bewertung der verfügbaren Unterstützung durch Werkzeuge.

5.2.3.1 Suchmaschinen

Die Nutzung von Suchmaschinen wie Google (www.google.org) stellt die frei verfügbare Standardmethode für die Informationsrecherche im Web dar. Die Vorgehensweise aller Suchmaschinen umfasst drei wesentliche Phasen:

- **Crawling**: In dieser Phase werden die verfügbaren Informationsquellen, in der Regel der öffentlich zugängliche Teil des Webs, durchsucht und hierfür ein Index aufgebaut, der ein effizientes Auffinden von Informationsquellen ermöglicht. Dieser Vorgang wird regelmäßig wiederholt, um die Aktualität des Index zu gewährleisten.
- **Search**: In dieser Phase haben Benutzer die Möglichkeit, Suchbegriffe und Themengruppen einzugeben. Die Suchmaschine nutzt diese Suchanfragen, um im aufgebauten Index relevante Informationsquellen zu identifizieren.
- **Ranking**: In der abschließenden Phase werden die gefundenen Informationsquellen mittels einer Ranking-Funktion bewertet. Das Ergebnis dieser Wertung entscheidet darüber, wie, insbesondere in welcher Reihenfolge, die Suchergebnisse für die Nutzer präsentiert werden.

Dieser Ansatz bezieht das gesamte öffentlich zugängliche Web ein und setzt zunächst keinerlei Expertenwissen bei den Nutzern voraus. Er ist somit grundsätzlich auch für die Suche im Bereich Life Cycle e-Valuation geeignet. Allerdings gibt es eine Reihe von Einschränkungen, die eine Verfeinerung des Ansatzes notwendig machen.

Hier ist zunächst die Qualität der zu erwartenden Suchergebnisse zu berücksichtigen. Bei jeder Art von Suche müssen hier insbesondere zwei Kenngrößen betrachtet werden. Einerseits ist dies die Genauigkeit (engl. precision) eines Suchergebnisses. Sie ist definiert als der Anteil relevanter Suchergebnisse an der Gesamtheit der Suchergebnisse. Demgegenüber sagt die Vollständigkeit (engl. recall) aus, wie hoch der Anteil der bei einer Suche gefundenen relevanten Dokumente (bzw. Datensätze) an den relevanten Dokumenten des insgesamt zur Verfügung stehenden Datenbestandes ist. Von einfachen Suchmaschinen ist im Bereich Life Cycle e-Valuation weder eine hohe Genauigkeit noch eine Vollständigkeit des Suchergebnisses zu erwarten, da keine detaillierten Informationen zu Nachhaltigkeit, Produkten und Dienstleistungen in die Suche einfließen. Ebenso besteht keine Möglichkeit, das Ranking speziell an die Suche im Themenumfeld anzupassen.

Die Effizienz der Suche ist ein weiterer Aspekt, der hier zu berücksichtigen ist. Allgemeine Suchmaschinen durchsuchen das gesamte Web. Damit erfolgt weder beim Crawling noch bei der Suche auf den indizierten Daten (Search) eine Einschränkung auf nachhaltigkeitsrelevante Informationsquellen. Es wird somit ein Datenbestand durchsucht, von dem offensichtlich ein erheblicher Teil keinen Zusammenhang zu den Themen im Bereich Life Cycle e-Valuation aufweist.

In den folgenden Abschnitten werden Technologien vorgestellt, die es ermöglichen die Phasen Crawling und Search besser auf die Suchaufgaben im Bereich Life Cycle e-Valuation zuzuschneiden.

5.2.3.2 Focused Crawling

Der Ansatz des Focused Crawling konzentriert sich in erster Linie auf die Vollständigkeit der Suchergebnisse und deren effiziente Berechnung. Vollständigkeit bezieht sich hier nicht auf das Web insgesamt, sondern auf die zu einem bestimmten Themenkomplex relevanten Dokumente (Chakrabarti et al. 1999). Diese sollen möglichst vollständig in einem Index erfasst werden und somit eine möglichst optimale Unterstützung für diese Suche nach Informationen in diesem Themenkomplex erreicht werden.

Zwei wesentliche Modifikationen an der Vorgehensweise des Crawlers machen einen allgemeinen Crawler zu einem Focused Crawler:

- Es werden nicht zu allen im Web verfügbaren Dokumenten Informationen in den Index der Suchmaschine übergenommen, sondern nur für solche, die für den interessierenden Themenkomplex als relevant bewertet werden.
- Wie bei der Standardvorgehensweise startet die Suche auch hier von einer initialen Menge an Dokumenten. Diese werden im Folgenden auch als Trainingsdaten bezeichnet. Allerdings erfolgt die Verfolgung von Links anders als die bei einem einfachen Crawler der Fall ist. Für jedes Dokument findet eine Relevanzbewertung statt. Diese soll eine Aussage dazu liefern, inwieweit die in einem Dokument enthaltenen Links verfolgt werden sollten, um Informationen zum betrachteten Themenkomplex zu finden. Nur wenn einem Dokument eine gewisse Mindestrelevanz zugeordnet werden kann, werden dessen Links überhaupt verfolgt. Die Reihenfolge in der Links berücksichtigt werden richtet sich ebenfalls nach der Relevanzbewertung.

In Abbildung 5-2 sind die beschriebenen Modifikationen schematisch dargestellt. Bei der Standardvorgehensweise werden alle Dokumente gelesen und einige davon (in der Abbildung schraffiert dargestellt) als Suchergebnis für den vorgegebenen Themenkomplex identifiziert. Beim Focused Crawling werden dagegen einige der möglichen Pfade nicht weiterverfolgt, da die über diese Pfade erreichbaren Ressourcen als nicht relevant eingestuft werden.

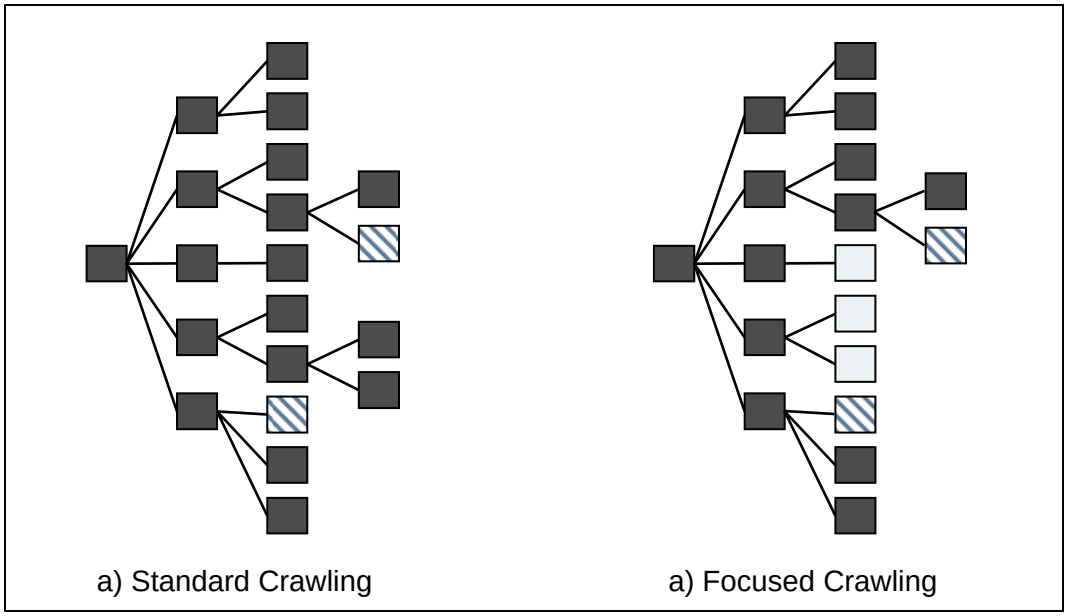

Abbildung 5-2: Crawling-Ansätze (nach Diligenti et al. 2000)

Somit ergeben sich für das Focused Crawling die beiden in Abbildung 5-3 dargestellten Verarbeitungszyklen (vgl. Ehrig und Maeche 2003). Zunächst geben die Benutzer vor, worin die Fokussierung des Crawlers bestehen soll. Mögliche Alternativen hierfür, wie Trainingsdaten oder eine Ontologie werden unten kurz erläutert. Diese Informationen werden dann im Zyklus 2 für den Aufbau des Index genutzt. Hierbei werden gefundene Dokumente zunächst vorverarbeitet, mit Hilfe einer Rankingfunktion bewertet und dann ggf. in den Index eingetragen. Dieser Index bildet im Zyklus 1 die Grundlage für die Suche der Benutzer. Abhängig von der Qualität der hier erzielten Ergebnisse können Benutzer den Fokus für das Crawling anpassen, indem sie z.B. andere, ergänzende Beispieldokumente bereitstellen.

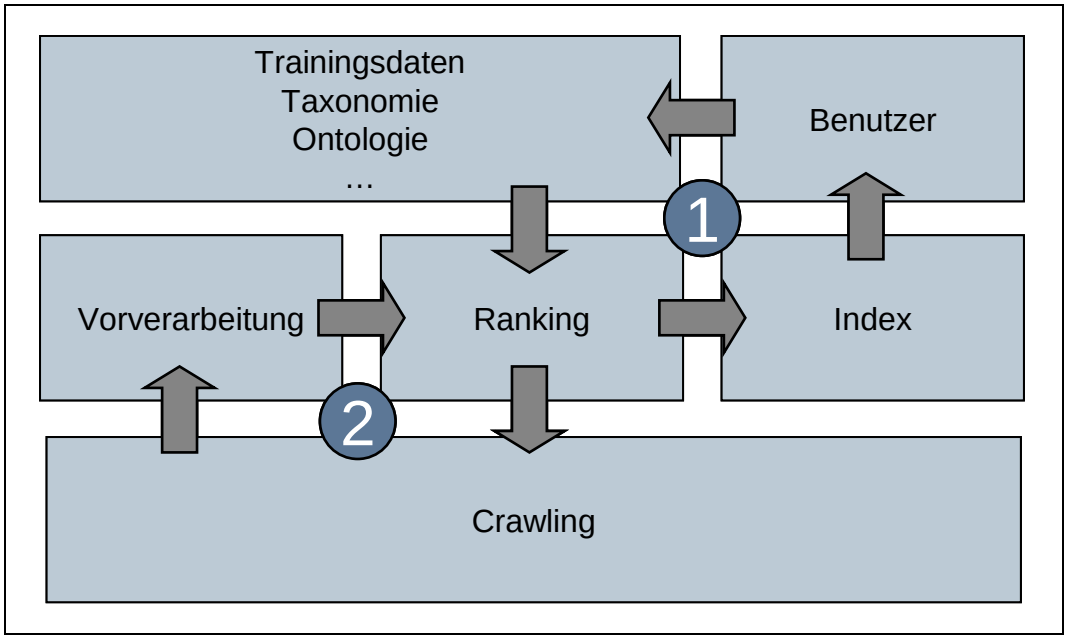

Abbildung 5-3: Verarbeitungszyklen beim Focused Crawling

Diese Vorgehensweise weist für Suchen, die sich auf einen bestimmten Kontext konzentrieren, mehrere Vorteile auf. Zunächst kann der Anteil der zu durchsuchenden Dokumente am gesamten Web reduziert werden. Dies spart einerseits Ressourcen und hat andererseits zur Folge, dass der aufgebaute Index in kürzeren Abständen aktualisiert werden kann. Darüber hinaus kann der Crawler durch die Bewertung der Relevanz einzelner Dokumente schneller zu den für den vorgegebenen Themenkomplex relevanten Ressourcen vordringen.

Diese Vorteile kommen auch bei einem Einsatz für den Bereich Life Cycle e-Valuation zum Tragen. Hier sind grundsätzlich nur Dokumente von Interesse, die einen Bezug zu Nachhaltigkeitsaspekten aufweisen. Werden relevante Themen (engl. topics) für diesen Bereich vorgegeben und ggf. Beispieldokumente bereitgestellt, dann kann ein Focused Crawler einen themen-spezifischen Index aufbauen. Die auf diesem Index aufsetzende Suchmaschine erlaubt dann einen schnellen Zugriff auf alle Informationen, die eine Nachhaltigkeitsbewertung unterstützen. Je nach Anwendungskontext kann auch eine zusätzliche Einschränkung auf ein bestimmtes Produkt, eine Dienstleistung oder eine Branche erfolgen. In jedem Fall bietet der Ansatz des Focused Crawling eine Möglichkeit die Suche gezielt auf einen Themenbereich zu fokussieren.

Bei der praktischen Umsetzung der Idee des Focused Crawling ergeben sich einige Probleme, die derzeit noch Gegenstand von Forschungsarbeiten sind. Die wichtigsten Aspekte sind:

- die optimale Reihenfolge, in der die in Dokumenten gefundenen Links verfolgt werden, sowie
- die Bewertung einzelner Dokumente bezüglich ihrer Relevanz für das vorgegebene Themengebiet (Ranking).

Für beide Aufgaben werden in den folgenden Abschnitten verschiedene Lösungsansätze exemplarisch erläutert.

Das erstgenannte Problem geht von der Beobachtung aus, dass im Web häufig ‚Inseln' von thematisch zusammenhängenden Dokumenten existieren. Zwischen solchen ‚Inseln' gibt es in der Regel nur Verbindungen über Seiten, die nicht demselben Themenkomplex zuzuordnen sind (vgl. Qin et al. 2004). Beim Crawling besteht somit die Gefahr, dass Dokumente in einer solchen ‚Insel' nur gefunden werden, wenn diese bereits durch die vorgegebenen Trainingsdaten repräsentiert wird.

Diligenti et al. beschreiben einen Ansatz, diesem Problem zu begegnen (Diligenti et al. 2000). Hierbei wird zu den vorgegebenen Trainingsdaten deren Kontext analysiert. Als Kontext werden in diesem Zusammenhang Web-Seiten verstanden, die auf das jeweilige Beispieldokument verweisen. Werden solche Links über mehrere Ebenen hinweg verfolgt, kann ein so genannter Kontext-Graph erstellt werden. Dieser Graph beschreibt unter anderem wie weit, d.h. wie viele Links, ein Dokument von einem Beispieldokument entfernt ist. Für diese Entfernung können automatisch Klassifikationsmodelle gebildet werden, die für ein beliebiges Dokument eine Aussage darüber erlauben, wieweit es voraussichtlich von einem relevanten Dokument entfernt ist. Diese Entfernung kann zur Steuerung des Crawling genutzt werden. Hierbei werden dann zunächst solche Links verfolgt, für die sich mit Hilfe der Klassifikation eine geringe Entfernung zu einem relevanten Dokument ergibt.

Bei einem anderen Ansatz werden die Links zwischen unterschiedlichen Themen analysiert. Grundlage hierfür ist eine Taxonomie sowie eine Reihe von Trainingsdaten für jeden Begriff in dieser Taxonomie. Beides wird dem Crawler vorgegeben. Die daraus abgeleitete Statistik wird genutzt um Regeln für die Link-Beziehungen zwischen den verschiedenen in der Taxonomie erfassten Themen aufzustellen. In solchen Regeln kann beispielsweise erfasst werden, dass ein Link von einer Seite zum Thema A auf eine Seite vom Thema B häufig wieder zu Informationen zu Thema A führt. Dieses Wissen erlaubt es, die Reihenfolge, in der Links beim Crawling verfolgt werden, zu steuern und damit die Lücke zwischen thematischen ‚Inseln' im Web zu überbrücken (Altingövde und Ulusoy 2004).

Focused Crawling beruht immer auf einer Bewertung eines Dokuments hinsichtlich dessen Relevanz für das vorgegebene Themengebiet. In der Regel wird hierzu auf der Grundlage von Trainingsdaten ein Klassifikationsmodell erstellt, das dann unverändert im kontinuierlichen Crawling-Prozess Verwendung findet. Es gibt allerdings auch Ansätze, die darauf abzielen, diese Klassifikation zu verbessern (vgl.

Sizov et al. 2003a, Sizov et al. 2003b und Chakrabarti et al. 2002). So kann das Klassifikationsmodell beispielsweise regelmäßig auf der Grundlage von als relevant identifizierten Dokumenten überarbeitet werden. In diesem Fall erweitert der Crawler den Satz der Trainingsdaten regelmäßig um weitere Dokumente, die als besonders relevant eingestuft wurden und verwendet diesen erweiterten Satz für eine erneute Berechnung des Klassifikationsmodells. In weiteren Arbeiten wird gezeigt, wie eine Ontologie genutzt werden kann, um die Relevanz eines Dokuments zu bestimmen (vgl. Ehrig und Maeche 2003) .

Diese kurze Übersicht zeigt, dass es für die zentralen Probleme beim Focused Crawling bereits eine Reihe viel versprechender Ansätze gibt. Deren Umsetzung erfolgte bisher aber meist in Form von Forschungsprototypen, kommerzielle Werkzeuge sind nicht verfügbar. Das System BINGO! Ist ein solcher Forschungsprototyp, der in ein umfassenderes und frei verfügbares System eingebettet ist. Dieses stellt somit einen möglichen Ausgangspunkt für die informationstechnische Unterstützung im Bereich Life Cycle e-Valuation dar.

5.2.3.3 Semantische Anfrageerweiterung

Werden thematische Einschränkungen nicht schon beim Crawling berücksichtigt, so besteht immer noch die Möglichkeit, dass die Suchmaschine die vom Benutzer eingegebene Anfrage modifiziert, d.h. in der Regel erweitert. Die primären Ziele sind: Die Erhöhung der Genauigkeit der gelieferten Ergebnismenge sowie die Erweiterung der Treffermenge, indem bei der Suche z.B. Synonyme mit berücksichtigt werden.

Alle Ansätze, die in diese Kategorie fallen gehen zunächst davon aus, dass der Benutzer seine Anfrage mit einigen wenigen Stichworten beschreibt und diese ohne weitere logische Verknüpfung aneinander reiht. Dies entspricht der überwiegenden Mehrheit der Suchanfragen, wie sie heute Google und andere Suchmaschinen entgegennehmen. Bei der Erweiterung der Anfrage durch die Suchmaschine stellt sich dann eine Reihe von Fragen:

- Kann die Anfrageerweiterung automatisch durch die Suchmaschine erfolgen oder werden dazu zusätzliche Informationen vom Benutzer abgefragt?
- Welche Modifikationen können an der Anfrage vorgenommen werden?
- Welche Informationen können für die Erweiterung der Anfrage genutzt werden?

Die automatische Anfrageerweiterung durch die Suchmaschine ist bereits seit einiger Zeit Forschungsgegenstand im Bereich Information Retrieval. Allerdings wird dort in der Regel von Anfragen ausgegangen, die eine Vielzahl von Begriffen (häufig zwischen 50 und 100) umfassen (vgl. de Lima und Pedersen 1999 sowie Hearst 1996). Bei solchen Anfragen ist es meist möglich, die gewünschte Bedeutung der Begriffe zu identifizieren und die Suchanfrage entsprechend zu modifizie-

ren und zu ergänzen. Bei Anfragen, die nur wenige Begriffe umfassen, ist dies in der Regel nicht möglich (vgl. Burton-Jones et al. 2003). In diesem Fall erscheint es Erfolg versprechender, zusätzliche Informationen vom Benutzer abzufragen.

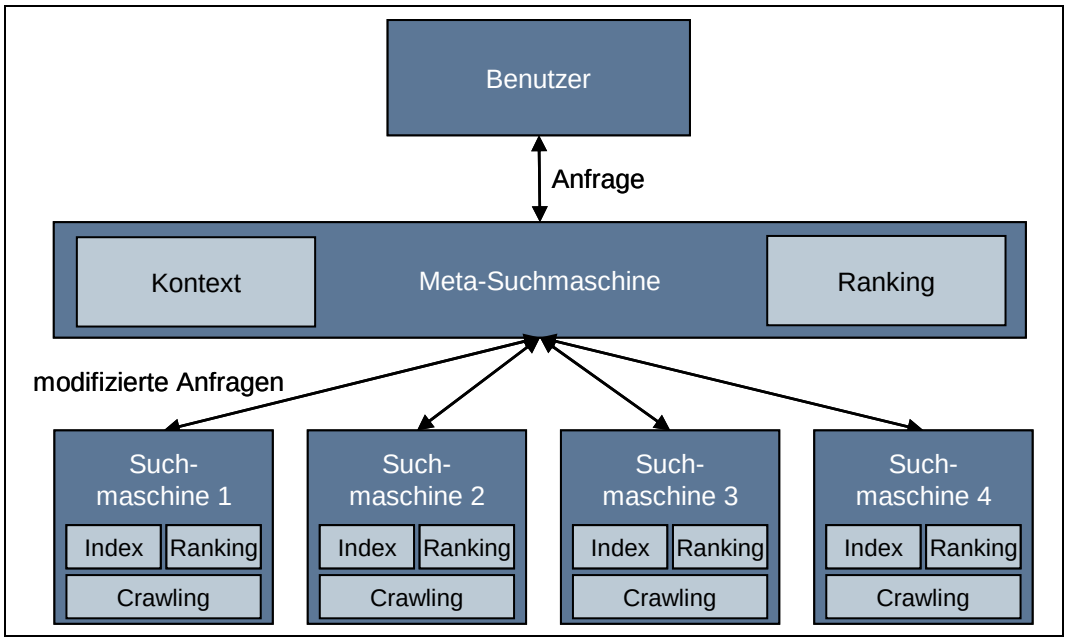

Abbildung 5-4: Meta-Suchmaschinen und Anfragemodifikation

Von Glover et al. wird ein Ansatz für eine Meta-Suchmaschine beschrieben. Abbildung 5-4 beschreibt den prinzipiellen Aufbau einer Meta-Suchmaschine. Diese verwaltet keinen eigenen Index, sondern greift bei einer Anfrage auf andere Suchmaschinen zu. Diese verwalten jeweils entsprechend ihren eigenen Index, deren Inhalt vor allem von der verfolgten Crawling-Strategie abhängt. Die Meta-Suchmaschine erhält damit von jeder Suchmaschine unterschiedliche Trefferlisten, deren Reihenfolge zudem von der lokal in einer Suchmaschine verwendeten Ranking-Funktion abhängt. Die Metasuchmaschine führt dann die einzelnen Ergebnismengen zu einer einzigen zusammen und bestimmt über eine eigene Rankingfunktion, in welcher Reihenfolge die Ergebnisse ausgegeben werden. Informationen, die von der Meta-Suchmaschine zur Modifikation der Anfragen und evtl. als Grundlage des Rankings verwendet werden, werden in Abbildung 5-4 als Kontext bezeichnet.

In der Suchanfrage an die Meta-Suchmaschine können beispielsweise zusätzlich zu den Suchbegriffen einzelne oder mehrere vom Benutzer ausgewählte Kategorien berücksichtigt werden (vgl. Glover et al. 2001). Diese Kategorien werden bei der Suche dann in unterschiedlicher Art und Weise eingesetzt:

- Sie beeinflussen in der Meta-Suchmaschine die Auswahl der Suchmaschinen, an die eine Anfrage weitergegeben wird.
- Pro Kategorie ist eine Menge von Regeln verfügbar, die eine gezielte Modifikation der Anfrage ermöglichen.
- Die Bewertung der Suchergebnisse der einzelnen Suchmaschinen wird ebenfalls durch die Kategorie beeinflusst.

Andere Ansätze ermöglichen die Auswahl der für die Suche gewünschten Kategorie auf der Basis einer Ontologie bzw. auf der Grundlage des Semantic Web (vgl. Guha et al. 2003). Hierbei werden die vom Benutzer eingegebenen Suchbegriffe analysiert und eine Menge in Frage kommender Kategorien identifiziert. Der Benutzer kann dann die gewünschte Kategorie auswählen, was es der Suchmaschine ermöglicht, gezielt nach Treffern für diese Kategorie zu suchen. Diese Vorgehensweise kann direkt umgesetzt werden, wenn die Zahl der zur Auswahl stehenden Kategorien begrenzt ist. Sobald eine Vielzahl von Kategorien identifiziert wird ist es nicht mehr sinnvoll, diese dem Benutzer vollständig zur Auswahl anzubieten. Dies kann beispielsweise der Fall sein, wenn nach einer bestimmten Person gesucht wird, die einen sehr häufig vertretenen Namen hat. Jedes Individuum mit diesem Namen stellt dann eine eigene Kategorie dar. Von Guha et al. wird für dieses Szenario vorgeschlagen, zunächst alle Suchergebnisse ohne semantische Erweiterung zu suchen und diese dem Benutzer anzuzeigen. Der Benutzer wählt dann ein Element der Trefferliste, das sich auf die gesuchte Kategorie bezieht aus und die Suchmaschine hat somit die Möglichkeit, die Suche weiter auf diese Kategorie einzuschränken.

In den Arbeiten zur semantischen Anfrageerweiterung werden unterschiedliche Möglichkeiten der Anfragemodifikation betrachtet, die in der Regel kombiniert eingesetzt werden sollten. Die wichtigsten Modifikationen sind:

- Bei einer Meta-Suchmaschine können spezifische Such-Optionen der genutzten Suchmaschinen gezielt eingesetzt werden. Solche Optionen erlauben es beispielsweise nach möglichst neuen Treffern zu suchen.
- Es können zusätzliche Einschränkungen in die Anfrage eingebracht werden. Dies ist z.B. sinnvoll wenn sichergestellt werden soll, dass mehrere Suchbegriffe in unmittelbarem Zusammenhang in einem gefundenen Dokument auftreten.
- Eine Anfrage kann um zusätzliche Begriffe ergänzt werden. Damit können z.B. zusätzliche Suchbegriffe eingebracht werden, die eine Einschränkung auf die gesuchte Kategorie zur folge haben oder aber es werden Oberbegriffe zu den vom Benutzer vorgegebenen Begriffen verwendet.
- Die Modifikation der Anfrage kann auch so erfolgen dass Begriffe ergänzt werden, die in den gesuchten Dokumenten nicht auftreten sollen.

Bleibt als weitere wichtige Frage, welche Kontextinformation als Grundlage für die Erweiterung von Anfragen dienen kann. In Burton-Jones et al. wird hier zwischen

lokalem und globalem Kontext unterschieden. Zum lokalen Kontext gehören hier alle Informationen, die direkt vom jeweiligen Benutzer vorgegeben werden. Hierzu gehören Beispieldokumente und Ontologien sowie Informationen zu den von diesem Benutzer bereits durchgeführten Suchen. Der globalen Kontext sind alle global zugänglichen Informationen zuzuordnen, die einen Beitrag zur Anfrageerweiterung leisten können. Hierzu gehören beispielsweise Thesauri (WordNet) und Sammlungen themen-spezifischer Ontologien.

Für den Bereich Life Cycle e-Valuation lässt sich der Ansatz der semantischen Anfrageerweiterung dann nutzen, wenn themen-spezifische Ontologien zur Nachhaltigkeitsbewertung verfügbar sind. Diese können als lokaler oder globaler Kontext in die Suchanfrage einfließen und erlauben es die Präzision der Suchergebnisse zu erhöhen.

5.3 Lead User Integration

5.3.1 Überblick über das Themenfeld

Zwei treibende Kräfte stellen die Basis für Produkt- und Serviceinnovationen dar: zum einen das Unternehmen selbst, das seine evtl. vorhandene Vorreiterrolle im jeweiligen Marktsegment behaupten oder erlange möchte. Der Innovationsprozess folgt in diesem Fall dem Prinzip des „Technology Push“ (Springer 2004), d.h. aus dem Unternehmen heraus werden Innovationen auf den Markt gebracht, wobei eine besondere Nachfrage seitens des Marktes nicht notwendigerweise vorliegt.
Im zweiten Fall hat eine Innovation ihren Ursprung im Markt selbst, d.h. das Unternehmen reagiert auf die Bedürfnisse des Marktes und entwickelt ein innovatives Produkt bzw. eine Dienstleistung, um diese Bedürfnisse zu befriedigen. Hier spricht man dann vom Prinzip des „Market Pull“ (Springer 2004). Beide Prozesse können jedoch nicht als vollständig gegeneinander isoliert betrachtet werden. Beispielsweise finden auch dann Marktbeobachtungen und -analysen statt, wenn die Innovation nicht den Forderungen des aktuellen Marktes entspringt, d.h. nicht im obigen Sinne direkt aus dem Unternehmen heraus initiiert wird. Im Gegenzug wird ein Unternehmen meist versuchen, einer vom Markt geforderten Innovation noch zusätzliche innovative Eigenschaften hinzuzufügen, um sich dadurch von der Konkurrenz abzuheben.

Mit der Lead-User-Methode (vgl. Hippel 1988) existiert ein Ansatz, beide Wege stärker zusammen zu führen und zu verzahnen. Auf Seiten des Marktes werden Personen identifiziert, die sich in besonderer Art und Weise durch spezielles Wissen oder ihre hohe Bereitschaft auszeichnen, innovative Ideen auszuprobieren und in ihrer Berufs- oder Alltagspraxis einzusetzen. Es kann sich dabei sowohl um fachspezifische Experten handeln, als auch um sog. Pionierkunden, im Englischen „Lead User“. Von solchen Pionierkunden verspricht man sich Anregungen zu Ideen, die aktuell vom Markt noch nicht in großem Umfang nachgefragt werden, bzw. deren fehlende Existenz aktuell nicht bewusst wahrgenommen wird. Mittel-

und längerfristig jedoch erhofft man sich für ihre Realisierung gute Absatzchancen. Zusammen mit diesen Personen werden dann im Rahmen von Workshops und Befragungen Ideen für neue Produkte oder Dienstleistungen entwickelt und konkretisiert. Im Idealfall schließt sich an diesen ersten Projektteil ein Praxisphase an, in der eine vorläufige Realisierung der Ideen zusammen mit den Lead Usern und evtl. weiteren Marktteilnehmern erprobt wird.

Die Lead-User-Thematik stellt den dritten Themenschwerpunkt des Projektes *nova-net* dar. Die nachfolgenden Erörterungen legen dar, inwiefern die Lead-User-Methode von der informationstechnischen Unterstützung profitieren kann, und wo diese anzusetzen hat.

5.3.2 Ziele der informationstechnischen Unterstützung im Themenfeld

Ziel der informationstechnischen Unterstützung des Lead-User-Verfahrens ist nicht seine methodische Weiterentwicklung, vielmehr ist der Schwerpunkt darin zu sehen, einzelne Phasen eines Lead-User-Projekes mit Mitteln der modernen IT zu unterstützen und so bisherige Schwächen des Verfahrens zu kompensieren. Ein Lead-User-Projekt gliedert sich in mehrere Phasen (vgl. Springer 2004) Angefangen von der Planung über die Identifikation von potenziellen Lead Usern, der Durchführung eines oder mehrerer Workshops bis hin zur Nachbereitung und Ergebnisauswertung. Diese Phasen wurden Im Rahmen des Projekts *nova-net* daraufhin untersucht, ob und in welchem Ausmaß eine informationstechnische Unterstützung sinnvoll ist und wo auf bereits existierende Standardsoftware zurückgegriffen werden kann bzw. wo neue Technologien und Werkzeuge entwickelt werden müssen (vgl. Springer 2004).

Ergebnis dieser Untersuchungen und darauf aufbauender Analysen ist die Erkenntnis, dass die frühen Phasen eines Lead-User-Projektes ein größeres noch unausgeschöpftes informationstechnisches Potenzial bieten. Während auf den Gebieten Projektplanung oder Workshopunterstützung und Ideengenerierung vielfältige Werkzeuge angeboten werden (vgl. beispielsweise 4.7.6), so ist insbesondere die Identifikation potenzieller Lead User ein Themengebiet, auf dem bisher wenig erreicht wurde. Aus diesem Grund erfolgt für die informationstechnische Unterstützung im Bereich Lead-User-Integration eine Fokussierung auf eben diese Phase der Identifikation.

Die zentrale Herausforderung in dieser frühen Phase eines Lead-User-Projektes ist es, in einem gegebenen Informationsraum Personen zu finden, die überdurchschnittliches Interesse an dem gewählten Thema zeigen oder sich durch besondere Kenntnisse bzw. Eigenschaften als kompetent auszeichnen. Dabei liegt die Schwierigkeit in der Beschaffenheit der zu durchsuchenden Informationsräume begründet. Handelt es sich um einfach strukturierte homogene Informationsräume wie z.B. Listen wissenschaftlicher Publikationen, so stellt sich diese Suche als nicht übermäßig kompliziert dar. Hier existieren in der Regel Kataloge mit Metada-

ten, anhand derer sich schnell und gezielt kompetente Personen ermitteln lassen. Anders verhält es sich hingegen, wenn komplexe heterogene Informationsräume wie z.B. das Internet durchsucht werden sollen. Insbesondere dann wird es kompliziert, wenn nicht nur ausgewiesene Experten (wovon im Fall der wissenschaftlichen Publikationen ausgegangen werden kann), sondern auch einfache Benutzer gefunden werden sollen, die sich lediglich dadurch auszeichnen, dass sie – wie auch immer geartete - Textbeiträge zu den untersuchten Themen veröffentlicht haben. Zu derartigen Beiträgen zählen neben eigenständigen Web-Seiten, beispielsweise einer privaten Homepage, auch Beiträge in Diskussionsforen oder Mailinglisten und News-Gruppen. In diesem Fall kann man nicht auf bekannte Strukturen wie dem formalen Aufbau einer wissenschaftlichen Veröffentlichung aufsetzen, sondern muss eine solche Struktur zuerst aus dem zugrunde liegenden Text ermitteln.

Auf letztgenannter Art von Suche soll im Folgenden der Schwerpunkt liegen. D.h. es soll erörtert werden, welche Technologien sinnvoll eingesetzt werden können, um o.g. Personen allgemein im Internet und im Besondern in seinen Anwendungen wie z.B. dem WWW, Mail oder Usenet zu finden.
Offensichtlich existieren hier starke Überlappungen mit den im Bereich „Lifecycle e-Valuation“ einsetzbaren Technologien, da auch hier die vorhandenen Informationsressourcen themenspezifisch durchsucht und verarbeitet werden müssen. An den entsprechenden Stellen wird daher auf die vorangegangenen Textabschnitte verwiesen.

5.3.3 Technologien zur informationstechnischen Unterstützung

Die Schwierigkeit bei der Identifikation von potenziellen Lead Usern besteht darin, dass sehr große Informationsräume durchsucht werden müssen. Zum einen müssen diese Informationsräume gefiltert werden hinsichtlich Relevanz für das betrachtete Themengebiet. Des Weiteren müssen dann Personen identifiziert werden, die mit den gefundenen Ressourcen in Verbindung stehen. Da dies aufgrund der Informationsfülle im Allgemeinen von Menschen allein nicht bewerkstelligt werden kann, ist hier die Unterstützung durch Computerprogramme gefordert. Sie können schnell und effizient auf große Datenmengen zugreifen und diese verarbeiten. Erreicht werden kann dies, wie bereits im vorangegangenen Kapitel über Life Cycle e-Valuation beschrieben, über sog. Crawler, die einen gegebenen Informationsraum mit gewissen Einschränkungen autonom durchsuchen können. Nachteil eines solchen reinen Crawlers ist, dass er per se nicht unterscheiden kann, welche gefundene Ressource tatsächlich für das Thema relevant ist, und welche nicht (vgl. 5.2.3.2). Für die Identifizierung potenzieller Lead User gilt es also zwei zentrale Probleme zu lösen:

1. Suchen und Auffinden möglichst großer Teilmengen der verfügbaren Textmaterialien zum jeweiligen Thema bei gleichzeitig hoher Relevanz der einzelnen Beiträge (Precision & Recall)

2. Extrahieren von Personeninformationen aus diesen Texten und Bewertung der Eignung dieser Personen als Lead User im behandelten Themenfeld

5.3.3.1 Bewertung der Relevanz von Texten

Das zentrale Problem bei der Verarbeitung großer Informationsmengen besteht darin, Wichtiges von Unwichtigem zu trennen bzw. eine Bewertung vorzunehmen, welche Informationen relevanter und damit wertvoller als andere sind. Ziel dabei ist, mit einer hohen Treffsicherheit bezüglich der Relevanz eine möglichst große Teilmenge der vorhandenen Informationen zu finden bzw. irrelevante Informationen herauszufiltern. Die zu dieser Aufgabe im Bereich Lifecycle e-Valuation vorgestellten Technologien rund um Ontologien und (semantische) Anfrageerweiterung können hier ebenfalls vorteilhaft eingesetzt werden.

Bei der oben beschriebenen Lifecycle e-Valuation jedoch sollen dem Benutzer primär eigenständige Dokumente präsentiert werden, die möglichst genau zum behandelten Thema passen. Der Mehrwert für den Nutzer ergibt sich daraus, dass ein großer Informationsraum systematisch eingeschränkt wird und somit der Zeitaufwand für die tatsächliche Informationsextraktion minimiert werden kann. Bei der Lead-User-Identifikation hingegen sind komplette Dokumente nicht notwendigerweise das Ziel. Vielmehr geht es darum, Personen zu finden, die sich qualifiziert zum jeweiligen Thema äußern. Bewertungsgrundlage müssen daher statt der kompletten Dokumente hier die evtl. in den Dokumenten enthaltenen Beiträge der einzelnen Personen sein. Grundlage der Bewertung mit den oben genannten Verfahren sind hier also bei Vorliegen solcher Beitragssammlungen vielmehr die einzelnen Beiträge selbst. Das Bewertungsverfahren kann mit diesen unterschiedlichen Eingaben dennoch prinzipiell auf die gleiche Art und Weise angewandt werden wie in Abschnitt 5.2.3 beschrieben.

Als weiteres Kriterium für die Relevanz kann betrachtet werden, von welchem Autor die jeweiligen Beiträge stammen bzw. auf welche anderen Beiträge sie sich beziehen. Ähnlich dem „PageRank“-Verfahren (Brin und Page 1998) können somit positiv bewertete Beiträge andere Beiträge in der Wertung positiv beeinflussen, wenn sie mit diesen in Verbindung stehen. Im Ergebnis führt dies dazu, dass aus der Relevanz der Beiträge eine Wertung für den jeweiligen Autor generiert werden kann, die eine Aussage darüber erlaubt, ob er zur manuellen Betrachtung selektiert werden soll oder nicht. Derartige Funktionalität wird beispielsweise von dem Werkzeug „Xpert-Finder“ (www.xpertfinder.de) angeboten. Hier kann der firmeninterne Mailverkehr analysiert und kategorisiert werden. Aus den beobachteten Kommunikationsstrukturen lassen sich dann Netzwerke aus Personen erstellen, die zu bestimmten Themen viel kommunizieren und daher eine gewisse Kompetenz vermuten lassen. Der Xpert-Finder konzentriert sich dabei allerdings im Wesentlichen auf firmeninterne Email- und News-Netzwerke.

5.3.3.2 Strukturanalyse von Dokumenten

Anhand der Beiträge in Diskussionen mit anderen Personen lässt sich die Kompetenz bzw. das Interesse einer Person am jeweiligen Themengebiet einschätzen. Diskussionsplattformen sind demnach eine essentielle Informationsquelle, um potenzielle Lead User zu identifizieren. Gängige Diskussionsplattformen (sog. Foren) sind in der Regel in Themen (Threads) unterteilt, welche sich wiederum aus den Beiträgen (Posts) der einzelnen Autoren zusammensetzen. Meist wird dabei ein Thread auf einer Seite dargestellt bzw. bei zunehmender Anzahl der Beiträge auch auf mehrere Seiten aufgeteilt.

Im Vergleich zu Dokumenten wie Informationsseiten für industrielle Produkte oder private Homepages müssen solche Forumsseiten unterschiedlich analysiert werden, da nicht mehr der gesamte Dokumenteninhalt einem Autor zugeordnet werden kann, sondern nun auch einzelne Dokumentenausschnitte eigenständige Informationseinheiten darstellen. Ist das Analysesystem mit einem Crawler (vgl. Abschnitt 5.2.3.2) gekoppelt, dann besteht die erste Schwierigkeit darin, das Vorliegen eines solchen Forums zu erkennen. Ist dieses Hindernis erfolgreich überwunden, so ergibt sich die nächste Schwierigkeit mit der automatischen Auftrennung des Gesamt-Threads in seine einzelnen Beitragsbestandteile. Für den menschlichen Betrachter ist dies ein Leichtes, da einzelnen Beiträge in der Regel so angeordnet sind, dass sie optisch voneinander unterschieden werden können. Analysesysteme hingegen besitzen diese visuellen Fähigkeiten nicht und müssen sich anderweitig behelfen.

Beide Probleme lassen sich durch Analyse der Seitenstruktur angehen. Wie in Abschnitt 4.6.2 dargelegt, ist das Standardformat zur Beschreibung einer Webseite HTML. Dieses besitzt jedoch, unter anderem, den Schwachpunkt, dass die Regeln für korrektes HTML lange Jahre nur mäßige Beachtung fanden und heute große Teile der verfügbaren HTML-Ressourcen aus fehlerhaftem Code bestehen. In der Folge haben Webbrowser häufig mehr oder weniger großen Interpretationsspielraum. Diese Uneindeutigkeit erschwert offensichtlich auch eine automatische Analyse der dem Dokument zugrunde liegenden Struktur. Um eine Analyse auf solchen potenziell schlecht oder fehlerhaft strukturierten Daten durchzuführen, bietet es sich an, die Daten zuerst in eine korrekte Struktur zu überführen. Dass diese Überführung korrekt verläuft kann dabei offensichtlich nicht garantiert werden, jedoch existieren Heuristiken, die die am weitesten verbreiteten Fehler identifizieren und korrigieren können. Solche Heuristiken werden von jeder HTML verarbeitenden Komponente (z.B. Browser) eingesetzt um eine regelmäßigere Verarbeitungsgrundlage zu erhalten. Ergebnis solch einer Normalisierung kann ein korrektes XHTML/XML-Dokument sein, dessen Struktur mit effizienten Algorithmen analysiert werden kann. Ein solcher Analyseansatz wären die Suche und der Vergleich von Teilstrukturen des hierarchisch organisierten XML-Dokuments, um Regelmäßigkeiten innerhalb der Struktur zu finden (vgl. Schlieder 2000). Zum einen lassen sich die einzelnen Beiträge so separieren, zum anderen kann dieses Verfahren auch auf die extrahierten Beiträge selbst angewandt werden, um weitere Informa-

tionen über Autor, Datum etc. zu erlangen. Letzteres gilt allerdings nur, wenn auch die einzelnen Beiträge wiederum einer regelmäßigen Form folgen, i.d.R. also nur, wenn es sich um maschinell erstellte Beitragssammlungen handelt (Beispiel: Foren).

Konnte die Struktur des Forums analysiert werden, so ist damit eine potenziell reiche Informationsquelle erschlossen. Mit geringem Aufwand lassen sich nun sämtliche Threads des Forums durchsuchen, die Beiträge extrahieren und einzelnen Autoren zuordnen. Durch die Thread-Struktur erhält man zusätzlich wertvolle Informationen darüber, wer mit wem über welches Thema diskutiert.

Um den Aufwand für das Crawling zu reduzieren, bietet es sich an, bei Foren oder komplexeren Webauftritten eine oftmals dort integrierte Suchfunktionalität zu nutzen. Mittels der im Abschnitt über Life Cycle e-Valuation angeführten semantischen Anfrageerweiterung lassen sich die Forenbeiträge vorfiltern und ein Großteil des Analyseaufwandes für irrelevante Beiträge kann entfallen. Dies setzt allerdings voraus, dass eine solche Suchfunktion erkannt und korrekt eingesetzt wird. Das Erkennen bedeutet, dass charakteristische Suchfeldstrukturen (einfache Suche, sog. „erweiterte Suche“ oder „Expertensuche“) in der Dokumentenstruktur erkannt werden müssen und der korrekte Einsatz bezieht sich auf die von der jeweiligen Suchfunktion erwartete Anfragesyntax.

Als Erweiterung der Strukturanalyse einzelner Dokumente kann die Analyse der Linkstruktur zwischen mehreren Dokumenten gelten. Aus der Dichte des Link-Graphen lassen sich Rückschlüsse ziehen auf das Vorhandensein von Experte- oder Beitragsnetzen. Zudem zeugt ein stark verlinktes Beitragsnetz von einer gewissen Kompetenz und sorgfältigen Vorgehensweise der Autoren, da davon ausgegangen werden kann, dass sie sich beim Erstellen des Beitrags intensiv mit der behandelten Thematik beschäftigt haben.

5.3.3.3 Extraktion von Autoreninformationen

Zentral für die Identifikation von potenziellen Lead Usern ist die Fähigkeit, herauszufinden, wer der Autor eines Beitrags ist. Diese Aufgabe stellt bei unterschiedlichen Beitragsarten eine unterschiedlich große Herausforderung dar. Während bei News-Beiträgen der Autor relativ einfach über die Kopfdaten der zugrunde liegenden Email ermittelt werden kann, so ist die Situation ungleich schwieriger wenn es sich um Beiträge in Form vom ganzen Webseiten oder klassischen Forumseinträge handelt.

Im Rahmen der Anstrengungen zum sog. Semantic Web (vgl. Abschnitt 4.1.2 sowie Abschnitt 4.1.3) werden Ressourcen zunehmend mit Metainformationen ausgestattet, die ein maschinelles Verarbeiten fördern. Mit dem Dublin-Core-Standard (DC) wurde in Abschnitt 4.1.2 ein Metadaten-System beschrieben, das es auch in dieser speziellen Anwendung erlaubt, verschiedene Informationen aus den

zugrunde liegenden Daten automatisch zu extrahieren. Von Interesse sind hier vor allem Informationen über den Autor und den Publikationszeitpunkt eines Beitrags. DC-Daten können in Form von XML direkt in (X)HTML-Dokumente eingebettet und aus diesen ausgelesen werden. Sie erfordern somit keine zusätzliche Annotations-Infrastruktur (vgl. 4.6.3).

Nachteilig bei obigem Ansatz ist, dass dieser nur dann zum Erfolg führt, wenn entsprechende Metadaten aktiv von den Verfassern der jeweiligen Textressource hinzugefügt wurden. Das mangelnde Bewusstsein für die Problematik der automatischen Verarbeitung bzw. das fehlende Interesse der Autoren, eine solche zu unterstützen setzt diesem Ansatz damit relativ enge Grenzen. Das Annotieren einer Ressource verlangt zusätzlichen Aufwand und damit in aller Regel zusätzliche Kosten. Im unternehmerischen Umfeld kann das evtl. akzeptiert werden, wenn sich dieser Aufwand als Investition betrachten lässt, d.h. wenn zu erwarten ist, dass durch diesen Mehraufwand letztlich ein finanzieller Vorteil entsteht. Bei Privatpersonen ist diese Bereitschaft in aller Regel nicht gegeben. Zudem haben Privatpersonen kein gesteigertes Interesse daran, ihre Informationen maschinell verarbeitbar zu machen. Diese maschinelle Verarbeitung dient in erster Linie kommerziellen Interessen Dritter, zu denen die jeweilige Person im Allgemeinen keinerlei Verbindung hat.

Liegen keine expliziten Zusatzinformationen in Form der genannten Metadaten über den Text vor, so bleibt der aufwändige Weg der Textanalyse. Diese lässt sich mit unterschiedlich hohem Aufwand betreiben, was naturgemäß zu unterschiedlich guten Resultaten führt.

Ein relativ einfacher und intuitiver Ansatz besteht darin, im Text nach Stichwörtern zu suchen, die auf das Vorhandensein der gewünschten Information schließen lassen, und dann die nähere Umgebung der Fundstelle intensiver zu analysieren. Beispielsweise lassen die Wörter „Mit freundlichen Grüßen“ darauf schließen, dass es sich um einen Brief handelt und als nächstes der Name des Verfassers genannt wird. Ähnliche Heuristiken lassen sich für viele andere Dokumententypen erstellen und führen trotz des relativ geringen Analyseaufwandes zu guten Ergebnissen (vgl. Baeza-Yates et al. 1999).

Weitere Ansätze zur Informationsextraktion aus Dokumenten entspringen der Computerlinguistik, insbesondere der Analyse von natürlichsprachlichen Texten (Natural Language Processing, NLP). Die Schwierigkeit bei der Verarbeitung natürlichsprachlicher Texte besteht darin, dass Worte bzw. Wortkombinationen ihre Bedeutung oft erst durch den Kontext erhalten, in den sie eingebettet sind. Demzufolge ist das gezielte Auslesen von spezifischen Informationen aus solchen Texten eine große Herausforderung für automatisierte Verfahren, da das Problem des algorithmischen Textverstehens noch nicht in dem Maße gelöst ist, dass beliebige Texte analysiert werden können.

Die hier behandelte Problemstellung hat jedoch nicht zur Aufgabe, einen natürlichsprachlichen Text komplette zu „verstehen“, es geht vielmehr darum einzelne relevante Informationseinheiten zu extrahieren, insbesondere Autoren- und Datumsinformationen. Es genügt hierzu also, das gestellte Problem soweit zu vereinfachen, dass zwar immer noch natürlichsprachlicher Text verarbeitet werden muss, es jedoch nicht mehr notwendig ist, diesen vollständig zu analysieren und zu verstehen. Eine solche Vorgehensweise ist die „Named Entity Recognition“ (NER). Unter dem Stichwort NER werden Methoden zusammengefasst, die aus natürlichsprachlichen Texten begriffliche Einheiten wie z.B. Orts-, Datums- oder Personenangaben zu extrahieren suchen. Sie stellen somit eine Untermenge der NLP-Verfahren dar.

Die Strategie bei NER ist, den Text in Bausteine aus Worten bzw. Wortkombinationen zu zerlegen und diese Bausteine mittels verschiedener Heuristiken zu klassifizieren. Dabei kommen zum einen Listen mit bekannten Begriffen zum tragen (beispielsweise Eigennamen), zum anderen können grammatikalische Regeln angewandt werden, wenn Teile eines Satzes bereits erkannt wurden (Bsp.: „Person A geht nach B“: wurde „A“ als Person identifiziert und „weiß“ der Algorithmus, dass „gehen“ eine Bewegung vom Start- zum Zielpunkt impliziert, dann kann er darauf schließen, dass „B“ mit einer gewissen Wahrscheinlichkeit eine Ortsangabe ist). Die oben genannte Stichwortsuche kann als eine weitere Komponente hier ebenfalls eingesetzt werden. Die Art der so erkennbaren Entitäten ist dabei stark abhängig vom verwendeten Algorithmus. Zudem sind diese Verfahren meist sehr sprachspezifisch, da sich die Regeln der Wortbildung und Grammatik von Sprache zu Sprache teilweise stark unterscheiden. Beispiele für existierende Softwarepakete zur NER sind „LingPipe“ (www.alias-i.com/lingpipe) oder „GATE“ (gate.ac.uk). Mittels NER lassen sich allerdings lediglich Entitäten innerhalb eines Textes ermitteln. Aussagen zu weitergehende Informationen, wie die im vorliegenden Fall benötigten Angaben, wer nun der Autor des Textes ist, sind von diesen Verfahren nicht zu erwarten. Hierfür wäre wiederum ein tieferes Verständnis des zugrunde liegenden Textes sowie dessen Kontexts notwendig.

5.3.3.4 Automatisiertes Lernen

Die Vielfältigkeit der möglichen Fragestellungen und Informationsbasen erlauben keine starren Systeme, die, einmal entwickelt, stets nach denselben Mustern suchen und aus deren Auftreten ihre Erkenntnisse ableiten. Neue Fragestellungen in neuen Themenbereichen erfordern stets eine Anpassung des Suchverhaltens sowie der Bewertungsfunktionen. Nicht zuletzt weil Individuen der jeweiligen Zielgruppe sich in ihrem Schreib- und Publikationsverhalten deutlich unterscheiden könne. Diese Vielfalt erfordert vielmehr eine dynamische Anpassung an die jeweiligen Gegebenheiten. Das schließt die Sprache ebenso ein wie Struktur und Gliederung der jeweiligen Informationsräume. Der Mensch realisiert eine solche Anpassung durch seine Fähigkeit, Verhaltensregeln und Methoden zu lernen. Das beinhaltet sowohl das Lernen aus Erfolgen, als auch das Lernen aus Misserfolgen.

Auf die hier behandelte Problemstellung übertragen bedeutet das, dass es wünschenswert wäre, die eingesetzten Algorithmen zu erweitern und ihnen in gewissem Umfang die Eigenschaft der Lernfähigkeit hinzuzufügen. Man erreicht damit eine höhere Flexibilität des Systems hinsichtlich wechselnder Benutzeranforderungen und eine stärkere Anpassbarkeit an die zu durchsuchenden Informationsräume. Insbesondere in zwei der oben genannten Bereiche ist dies vorteilhaft:

- Anpassung der Bewertungsfunktion zum Filtern von Beiträgen
  Die Suche nach relevanten Beiträgen (beispielsweise beim Focused Crawling, vgl. Abschnitt 5.2.3.2) beginn mit einer initialen Bewertungsfunktion. Mit Fortschreiten der Suche wächst die Menge der bereits als relevant/irrelevant bewerteten Beiträge an und stellt somit eine ergiebige Lernquelle dar. Die Bewertungsfunktion kann nun angepasst werden und anhand der als relevant eingestuften Beiträge erlernen, welche weiteren Eigenschaften „offensichtlich relevante“ Beiträge haben (analog kann das Lernen aus Negativbeispielen zu besseren Filterergebnissen führen). Dieses neu erlangte „Wissen“ kann dann wiederum auf weitere Beiträge angewandt werden. Im Verlauf der Suche können Beiträge damit immer sicherer als relevant/irrelevant erkannt werden und zudem werden mit einer höheren Wahrscheinlichkeit auch solche relevanten Beiträge gefunden, die von der initialen Bewertungsfunktion als irrelevant eingestuft würden, da diese z.B. mit der verwendeten Terminologie nicht umzugehen verstand.

- Erlernen von wiederkehrenden Dokumentenstrukturen
  Die Strukturanalyse von Dokumenten ist eine rechenzeitintensive Operation. Im Allgemeinen ist es jedoch nicht nötig, jedes Dokument einer vollständigen Analyse zu unterziehen, um herauszufinden, um was für eine Art Dokument es sich handelt. Vielmehr wird man versuchen, anhand einiger weniger, schnell zu bestimmender Merkmale, mit einer gewissen Wahrscheinlichkeit auf einen Dokumententyp zu schließen. Erst wenn sich diese Annahme als falsch herausgestellt hat, erfolgt dann eine komplette Analyse. Diese Problemstellung bietet sich für ein maschinelles Lernen an, da die hieraus gewonnenen Erkenntnisse bei zukünftigen Klassifizierungen die Verarbeitungszeit erheblich reduzieren können.
  Zieht man ferner in Betracht, dass beispielsweise Web-Foren in der Regel nicht von jedem Anbieter eigenständig implementiert werden, sondern ein Großteil aller Foren auf wenigen kommerziell oder frei verfügbaren Systemen aufbauen, so lassen sich hier die Erkenntnisse, die man aus der Analyse eines Forums gewonnen hat, rechenzeitsparend auf andere Foren übertragen. Meist unterscheiden sich Foren optisch zwar auch dann von einander, wenn sie auf demselben Foren-System aufbauen, in ihrer Struktur jedoch gleichen sie sich meist bzw. sind sogar identisch. Hat die Analysefunktion eine solche Forenstruktur einmal erlernt, so bleibt als einzige Schwierigkeit, die der Wiedererkennung. Ist diese auch nicht trivial, so erfordert sie in aller Regel dennoch weniger Aufwand als eine vollständige Strukturanalyse (vgl. Abschnitt 5.3.3.2).

Insbesondere im Bereich der Klassifikation und Relevanzbewertung von Texten existieren vielfältige Techniken und Methoden bzw. werden derzeit erforscht. Beispielsweise basieren die sog. Bayes-Filter auf Erkenntnissen aus der Wahrscheinlichkeitstheorie (Bayes-Theorem). Mittels Listen von Worthäufigkeiten in Positiv- und Negativbeispielen versucht der Bayes-Filter, neue Texte ebenfalls als relevant/irrelevant zu klassifizieren. Die derzeit wohl bekanntesten Anwendungsgebiete solcher Filter liegen in der Spam-Erkennung. Weit verbreitet ist hier der Filter „SpamAssassin“ (spamassassin.apache.org).

Weitere Lernansätze werden z.B. mit künstlichen neuronalen Netzen (KNN, vgl. Rojas 1993) und sog. „Support Vector Machines“ (SVM, vgl. Boser 1992) verfolgt. Während es sich bei SVM um eine weitere statistisch fundierte Methode handelt, die im Wesentlichen auf dem Prinzip der Hyperebenentrennung basiert (siehe Wapnik 1979), so sind KNN einer äußerst leistungsfähigen Lernkomponente nachempfunden – dem Gehirn, wobei ihr Verhalten dem Zusammenspiel von Nervenzellen nachempfunden ist. Ihr Einsatz erfolgt meist im Rahmen von Aufgaben der Mustererkennung. Beispielhafte Implementierungen dieser Methoden finden sich in „SVMlight“ (svmlight.joachims.org) oder „SNNS“ (www-ra.informatik.uni-tuebingen.de/SNNS).

Lernalgorithmen lassen sich unterteilen in Selbstlernende und solche, die vom Benutzer aktiv trainiert werden müssen. Der einfachere Ansatz ist im Regelfall der des nutzertrainierbaren Algorithmus, nachteilig ist indes, dass eine regelmäßige Benutzerinteraktion notwendig ist und der Algorithmus nicht über längere Zeit hinweg eigenständig operieren und sich dabei verbessern kann.
Selbstlernende Algorithmen hingegen versuchen mittels an die Situation angepasster Heuristiken selbst herauszufinden, welche Textbeispiele als positiv bzw. negativ zu erachten sind und machen damit die Benutzerinteraktion überflüssig. Offensichtlich haben sie jedoch den Nachteil, dass sich Fehler einschleichen können und diese Fehler durch die Lernkomponente so verstärkt werden, dass die Bewertungsergebnisse letztlich nicht mehr zu gebrauchen sind, weil zu viele irrelevante Texte als relevant erkannt bzw. zu viele relevante Text als irrelevant eingestuft wurden.

5.3.3.5 Bereitstellung der Funktionalität und deren Integration in Anwendungen

Idealerweise wird die Such- und Analysefunktionalität nicht als eine eigenständige und in sich geschlossene Anwendung bereitgestellt, sondern als modulares Dienstepaket auf Basis einer Client-/Server-Struktur. Dies hat den Vorteil, dass die jeweils benötigte Funktionalität relativ einfach in unterschiedliche Anwendungen integriert werden kann und die Datenhaltung sowie die Datenpflege zentral mit geringem Aufwand durchgeführt werden können. Im derzeitigen Technologiekontext liegt eine Realisierung auf Basis von Web Services nahe. Web Services erlauben das Einbinden der angebotenen Funktionalität direkt in Clientanwendungen, oder aber als Komponente in weiteren Service-Anwendungen (vgl. Abschnitt 4.7.3). Sie

lassen sich unterschiedlichen Programmiersprachen und -umgebungen realisieren, so dass auch die Portabilität zwischen verschiedenen Betriebssystemen bzw. Hardwarearchitekturen ohne großen zusätzlichen Implementierungsaufwand erreicht werden kann.

5.4 Zusammenfassung

Die vorangehenden Abschnitte haben einerseits das Potenzial für die informationstechnische Unterstützung in den Themenfeldern Trendmonitoring im Szenario-Management, Life Cycle e-Valuation und Lead User Integration vorgestellt. Andererseits wurde deutlich, dass in keinem der Themenfelder eine umfassende informationstechnische Unterstützung durch existierende Software-Tools verfügbar ist.

Zur Unterstützung des Trendmonitoring im Szenario-Management existieren derzeit einzelne Werkzeuge, die Teilaspekte des Szenario-Managements abdecken. Eine umfassende Lösung, die einerseits den gesamten Szenario-Erstellungsprozess unterstützt und andererseits auch eine systematische Verfolgung von Änderungen in den erfassten Szenarien ermöglicht, existiert nicht. Wie gezeigt, können diese Defizite über ein so genanntes Szenario-Management-Framework adressiert werden. Dieses bildet einen flexiblen Rahmen für das unternehmens-spezifische Zusammenführen unterschiedlicher Szenario-Management-Funktionalität in benutzerdefinierten Workflows.

Für die Fragestellungen im Bereich Life Cycle e-Valuation existiert ebenfalls keine spezifische informationstechnische Unterstützung. Wie gezeigt, kann eine solche Unterstützung allerdings auf Technologien aufbauen, die entwickelt wurden, um wesentliche Defizite von Suchmaschinen zu beheben. Hierzu zählt einerseits der Ansatz des Focused Crawling, der es erlaubt, bereits beim Crawling nur solche Dokumente zu berücksichtigen, die für den vorgegebenen Themenkomplex relevant sind. Diese Auswahl kann beispielsweise auf der Grundlage themenspezifischer Ontologien zur Nachhaltigkeitsbewertung erfolgen. Solche Ontologien sind genauso für den zweiten relevanten Ansatz, die semantische Anfrageerweiterung von Bedeutung. Bei dieser ist es das Ziel, die Genauigkeit der gelieferten Ergebnismenge zu erhöhen, indem bei der Suche beispielsweise Kenntnisse über Synonyme eingesetzt und zur Modifikation der Suchanfrage genutzt werden.

Die zentrale Herausforderung bei der informationstechnischen Unterstützung eines Lead-User-Projektes ist es, in einem gegebenen Informationsraum Personen zu finden, die überdurchschnittliches Interesse an dem gewählten Thema zeigen oder sich durch besondere Kenntnisse bzw. Eigenschaften als kompetent auszeichnen. Dabei liegt die Schwierigkeit insbesondere in der Heterogenität und Beschaffenheit der zu durchsuchenden Informationsräume begründet. Die detaillierte Analyse verfügbarer Technologien hat gezeigt, dass für die wichtigsten Teilaspekte (Relevanzbewertung von Texten, Strukturanalyse von Dokumenten, Extraktion von Autoreninformationen, automatisiertes Lernen) die Basistechnologien für eine infor-

mationstechnische Unterstützung gegeben sind. Eine Kombination dieser Technologien mit dem Ziel, die spezifischen Fragestellungen im Themenfeld Lead User Integration zu adressieren, steht allerdings noch aus.

6 Anhang: Übersicht verfügbarer Software

In der folgenden Tabelle ist beispielhaft Software aufgeführt, die für die informationstechnische Unterstützung der drei Themenfelder eingesetzt werden kann. Es werden dabei sowohl prototypische Implementierungen, die im Rahmen von Forschungsprojekten entstanden sind berücksichtigt, als auch kommerzielle Software. Der Schwerpunkt der Zusammenstellung liegt auf Technologien, die in Kapitel 5 als für zumindest eines der Themenfelder als besonders wichtig identifiziert wurden.

Produkt	Hersteller	URL	Kurzbeschreibung	Zuordnung
CIM 8.0	Center for Futures Research, St. Gallen	http://www.sgzz.ch/index_research_software.php	Das Werkzeug CIM 8.0 unterstützt die Durchführung der Cross-Impact-Analyse im Rahmen von Workshops.	Trendmonitoring im Szenario-Management
INKA 3	Geschka & Partner	http://www.geschka.de/	Das Werkzeug INKA 3 stützt sich e-benfalls auf die Szenario-Technik und unterstützt die so genannte Konsis-tenz-Analyse.	Trendmonitoring im Szenario-Management
Szeno-Plan	SINUS Software Consulting GmbH	http://www.sinus-online.com/	Szeno-Plan baut ebenfalls auf die Szenario-Technik auf und unterstützt sowohl die Konsistenz-Analyse als auch die Cross-Impact-Analyse.	Trendmonitoring im Szenario-Management
Szenario-Manager	UNITY AG	http://www.szenario-manager.de	Der Szenario-Manager ist ein Werk-zeug, das die größte Funktionalität in der Reihe der Szenario-Technik-Anwendungen bietet. Die Szenarien basieren ebenfalls auf eine Menge von Einflussfaktoren sowie deren Ausprä-gungen.	Trendmonitoring im Szenario-Management
BINGO!	Max-Planck-Institut für Informatik , Saar-brücken	http://www.mpi-sb.mpg.de/units/ag5/software/bingo/	Forschungsprototyp, der die Idee des Focused Crawling umsetzt. Der Craw-ler wird auf der Basis von bereits klas-sifizierten Beispieldokumenten trai-niert. Bei der Bewertung von Doku-menten wird darüber hinaus eine On-tologie genutzt.	Life Cycle e-Valuation

Produkt	Hersteller	URL	Kurzbeschreibung	Zuordnung
Metis focused Crawler	Universität Karlsruhe	http://metis.ontoware.org/	Forschungsprototyp eines Focused Crawler, der anhand einer definierbaren Ontologie Suchergebnisse bewertet. Dabei werden sowohl evtl. vorhandene Metadaten, als auch natürlichsprachliche Elemente der Webseiten als Bewertungsgrundlage herangezogen.	Life Cycle e-Valuation, Lead User Identifikation
AskMe Enterprise	AskMe Corporation	http://www.askmecorp.com/	Kommerzielles Wissensmanagement-System, das unterschiedlichste Dienste rund um eine Wissensdatenbank anbietet. Die Nutzung vieler dieser Dienste kann per Email erfolgen. Einer dieser Dienste unterstützt auch die themen-spezifische Suche von Experten innerhalb eines Unternehmens.	Lead User Identifikation Expertensuche
SVMLight	T. Joachims, Universität Dortmund	http://svmlight.joachims.org/	Forschungsprototyp zur Textklassifikation und –Bewertung mittels Support-Vector-Machine-Technologie	Life Cycle e-Valuation, Lead User Identifikation
WebSPHINX	Carnegie Mellon University	http://www-2.cs.cmu.edu/~rcm/websphinx/	Java-Bibliothek zum Aufbau eines Web-Crawlers. Flexible API erlaubt das Einbinden eigener Bewertungsfunktionen und Heuristiken zur Beeinflussung der Suche.	Life Cycle e-Valuation, Lead User Identifikation

Produkt	Hersteller	URL	Kurzbeschreibung	Zuordnung
LingPipe	Alias-i, Inc.	http://www.alias-i.com/lingpipe	Nachfolger eines Forschungsprototyps zur Named Entity Recognition in einfachen Texten sowie HTML/XML-Dokumenten. Modell für die englische Sprache wird mitgeliefert, das System ist jedoch offen für weitere Sprachen.	Life Cycle e-Valuation, Lead User Identifikation
Xpertfinder	Fraunhofer IPA	http://www.xpertfinder.de/	Kommerzielles System zur Identifizierung von Experten und Expertennetzwerken durch Analyse der unternehmensinternen Kommunikation.	Lead User Identifikation

7 Literaturverzeichnis

Altingövde, I. S.; Ulusoy, Ö. (2004) *Exploiting Interclass Rules for Focused Crawling*. IEEE Intelligent Systems 19(6): 66-73

Baeza-Yates, R.; Ribeiro-Neto, B.; Navarro, G. (1999): *Query Languages*, In: Modern Information Retrieval. ACM-Press/Addison-Wesley

Beer, W.; Birngruber, D.; Mössenbäck, H.; Wöß, A. (2003): *Die .NET-Technologi. Grundlagen und Anwendungsprogrammierung.* dpunkt.verlag, Heidelberg

Boser, B.; Guyon, I.; Vapnik, V. (1992): *An training algorithm for optimal margin classifiers*, In: Fifth Annual Workshop on Computational Learning Theory, Seite 144-152, ACM, Pittsburgh

Brin, S.; Page, L. (1998): *The Anatomy of a Large-Scale Hypertextual Web Search Engine*. In: Proceedings of the Seventh International Web Conference (WWW98), 1998

Bullinger, H.-J. ; Bucher, M. ; Kretschmann, T.; Müller, M.: Knowledge meets System- wissensbasierte Informationssysteme, Fraunhofer IRB Verlag, Stuttgart, 2001

Burton-Jones, A.; Veda C. Storey, Vijayan Sugumaran, Sandeep Purao (2003) A Heuristic-Based Methodology for Semantic Augmentation of User Queries on the Web. ER 2003: 476-489

Chakrabarti, S.; van den Berg, M.; Dom, B. (1999): Focused Crawling: a new approach to topic-specific Web resource discovery. Proc. of the 8th International World-Wide Web Conference (WWW8), Toronto, Canada, May 11-14, 1999

Chakrabarti, S.; Punera, K.; Subramanyam, M.: Accelerated Focused Crawling through Online Relevance Feedback (2002): Proceedings of the Eleventh International World Wide Web Conference, WWW2002, Honolulu, Hawaii, USA, 7-11 May 2002

CIM (2003): *Cross Impact Model*. http://www.sgzz.ch/index_research_software.php

Davis, J. R.; Lagoze, C. (2000): *NCSTRL: Design and deployment of globally distributed digital library.* In: Journal of the American Society of Information Science 51 Nr. 3 Seite 273-280

Diligenti, M.; Coetzee, F. M.; Lawrence, S.; Giles, C. L.; Gori, M. (2000): Focused Crawling Using Context Graphs. International Conference on Very Large Databases, VLDB 2000, Cairo, Egypt, 2000

Dinter, B.; Sapia, C.; Höfling, G.; Blaschka, M. (1998): *The OLAP Market: State of the Art and Research Issues. In: Proceedings of the first International Workshop on Data Warehousing and OLAP (DOLAP),* Bethesda, Maryland, USA

Ehrig, M.; Maedche, A. (2003) Ontology-Focused Crawling of Web Documents. In: Proceedings of the 2003 ACM Symposium on Applied Computing (SAC), March 9-12, 2003, Melbourne, FL, USA

Elmasri, R.; Navathe, S. B. (2002): *Grundlagen von Datenbanksystemen.* Pearson Studium, München

Engelhardt, K. (2002): *Dokumenten-Management-Systeme : Marktübersicht, Hersteller, Produkte.* Verband Organisations- und Informationssysteme e.V., Darmstadt

Fensel, D.; van Harmelen, F.; Horrocks, Ian (2002): *OIL and DAML+OIL: Ontology Languages for the Semantic Web*, in *Towards the Semantic Web – Ontology-Driven Knowledge Management*. J.Wiley, New York

Ferber, R. (2003): *Information Retrieval.* dpunkt.verlag, Heidelberg

Fichter, K.; Kiehne, D. O. (2004): *Trendmonitoring im Szenario-Management.* Forschungsprojekt nova-net, Stuttgart

Gausemeier, J.; Fink, A.; Schlake, O. (1996): *Szenario-Management: Planen und Führen mit Szenarien.* Carl Hanser Verlag, München

Glover, E. J.; Lawrence, S.; Gordon, M. D.; Birmingham, W. P.; Giles, C. L. (2001) Web search – your way. Communications of the ACM, 44(12):97-102, 12 2001

Götzer, K.; Schneiderath, U.;Maier, B.; Komke, T. (2004): *Dokumenten-Management.* dpunkt.verlag, Heidleberg

Guha, R.; McCool, R.; Miller, E. (2003) Semantic search. In: Proceedings of the Twelfth International World Wide Web Conference, WWW2003, Budapest, Hungary, 20-24 May 2003

Haase O. (2001): *Kommunikation in verteilten Anwendungen*, R. Oldenbourg Verlag, München Wien

Han, J.; Kamber, M. (2001): *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers, San Francisco

Hauschildt, J. (1997): *Innovationsmanagement.* Vahlen, München

Hearst, M. A. (1996) Improving Full-Text Precision on Short Queries using Simple Constraints. In: Fifth Annual Symposium on Document Analysis and Information Retrieval April 15 - 17, 1996, Las Vegas, Nevada, USA

Hippel, Eric (1988): *The Source of Innovation*, Oxford University Press, New York, Oxford

Hübner, H. (2002): *Integratives Innovationsmanagement.* Erich Schmidt Verlag, Berlin

INKA: INKA 3. http://www.geschka.de

International Organization for Standardization and International Electrotechnical Committee (1984): *Open Systems Interconnection: Basic Reference Model.* ISO 7498

ISO (2003): ISO/IEC 9075-14:2003, *Information technology -- Database languages -- SQL -- Part 14: XML-Related Specifications (SQL/XML)*

Klingenhöller, H. (2001): *Dokumentenmanagementsysteme: Handbuch zur Einführung.* Springer Verlag, Heidelberg

Lang, C.; Springer, S.; Beucker, S. (2004): *Life Cycle e-Valuation Produkt, Service, System.* Forschungsprojekt nova-net, Stuttgart de Lima, E. F.; Pedersen, J. O. (1999) Phrase Recognition and Expansion for Short, Precision-biased Queries based on a Query Log. In 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA

Melton, J.; Simon, A. (2002): SQL:1999: *Understanding Relational Language Components.* Academic Press, San Francisco

Melton, J. (2003): Advanced SQL:1999: *Understanding Object-Relational and Other Advanced Features.* Elsevier Science, San Francisco

OASIS (2003): *Universal Description, Discovery and Integration (UDDI) Version 3.0.1*, http://uddi.org/pubs/uddi_v3.htm

Peritsch, M. (2000): *Wissensbasiertes Innovationsmanagement – Analyse, Gestaltung, Implementierung.* Deutscher Universitäts-Verlag GmbH, Wiesbaden

Pölzl, A. (2002): *Umweltorientiertes Innovationsmanagement – Eine theoretische und empirische Analyse.* Verlag Wissenschaft & Praxis, Sternenfels

Qin, J.; Zhou, Y.; Chau, M. (2004) Building domain-specific web collections for scientific digital libraries: a meta-search enhanced focused crawling method. Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, Tuscon, AZ, USA, June 7-11

Reibnitz, U. (1991): Szenario-Technik: Instrumente für die unternehmerische und persönliche Erfolgsplanung, Gabler Verlag, Wiesbaden

RFC 2413 (1998): *RFC 2413 - Dublin Core Metadata for Resource Discovery,* Network Working Group, http://www.faqs.org/rfcs/rfc2413.html

Rojas, R. (1993): Theorie der neuronalen Netze, Springer-Verlag, Berlin

Rose, M. (1990): *The Open Book: A Practical Perspective on OSI.* Prentice Hall, Englewood Cliffs, New Jersey

Rothfuss, G.; Ried, C. (2003): *Content Management mit XML.* Springer-Verlag, Berlin, Heidelberg

Schlieder, T (2000).: *Strukturelle Ähnlichkeitssuch in XML-Dokumenten*. In: 12. GI-Workshop "Grundlagen von Datenbanken", Plön

Sizov, S.; Biwer, M.; Graupmann, J.; Siersdorfer, S.; Theobald, M.; Weikum, G.; Zimmer, P. (2003a) The BINGO! System for Information Portal Generation and Expert Web Search. The 1st Semiannual Conference on Innovative Data Systems Research (CIDR), Asilomar(CA)

Sizov, S.; Graupmann, J.; Theobald, M. (2003b) From Focused Crawling to Expert Information: An Application Framework for Web Exploration and Portal Generation. In: Proceedings of 29th International Conference on Very Large Data Bases, September 9-12, 2003

Spath, D.; Ardilio, A.; Auernhammer, K.; Kohn, S. (2004): *Marktstudie Innovationssysteme – IT Unterstützung im Innovationsmanagement.* Fraunhofer IRB Verlag, Stuttgart

Springer, S.; Beucker, S.; Lang, C.; Bierter, W. (2004): *Lead User Integration.* Forschungsprojekt nova-net, Stuttgart

Sun Microsystems (2003): *Java 2 Platform, Enterprise Edition (J2EE 1.4)*, http://java.sun.com/j2ee/

SzenMan (2004): Szenario-Manager. http://www.szenario-manager.de

Szeno-Plan (2004): Szeno-Plan. SINUS Software and Consulting GmbH

W3C (1999) : *RDF Model and Syntax Specification*, http://www.w3.org/TR/rdf-primer

W3C (2001a) : *Web Services Description Language (WSDL) 1.1*, http://www.w3.org/TR/wsdl

W3C (2001b): *Annotea*, http://www.w3.org/2001/Annotea

W3C (2001c): Synchronized Multimedia Integration Language (SMIL 2.0), http://w3.org/TR/smil20/

W3C (2002) : *Web Services Activity*, http://www.w3.org/2002/ws/

W3C (2003): *SOAP Version 1.2*, http://www.w3.org/TR/soap/

W3C (2004a): *OWL Web Ontology Language Reference*, http://www.w3.org/TR/owl-ref

W3C (2004b): *XQuery 1.0: An XML Query Language*. W3C Working Draft 23 July 2004, http://www.w3.org/TR/2004/WD-xquery-20040723/

Wapnik, W.; Tscherwonenkis, A.: *Theorie der Zeichenerkennung*, Akademie-Verlag, Berlin, 1979

Wilhelm, S; Spath, D.: Information und Kommunikation in der Produktion – Ergebnisse einer Unternehmensbefragung. Handlungsbedarf und Handlungsempfehlungen für die Informationsbewirtschaftung direkt produktiver Bereiche, Fraunhofer IRB Verlag, Stuttgart, 2003

Wöhr, H. (2004): *Web Technologien*. dpunkt.verlag, Heidelberg

Zschau, O.; Traub, D.; Zahradka, R. (2002): *Web Content Management.* Galileo Press GmbH, Bonn