Die Infrastruktur und das Dokumentenmanagement von XML-basierten wissenschaftlichen Informationssystemen

Eckhardt Schön

Technische Universität Ilmenau, Fakultät für Informatik und Automatisierung,
Institut für Praktische Informatik und Medieninformatik

Abstract. Dieser Vortrag soll die Möglichkeiten aufzeigen, die die Extensible Markup Language für den Aufbau effizienter Informationssysteme bietet. Aus verschiedenen Gründen dürften sich XML-basierte Systeme zuerst im wissenschaftlichen Umfeld durchsetzen. Deshalb wird auf diesen Bereich genauer eingegangen.

1. XML als Dokumentenformat für Informationssysteme

Ein Informationssystem unterscheidet sich von einer Datenbank u.a. darin, dass die Daten dem Nutzer in einer Form präsentiert werden, die es ihm einfach macht, diese Daten zu rezipieren und seinem Wissen hinzuzufügen. Es muss folgende Aufgaben unterstützen, das Sammeln, Importieren, Speichern, Indexieren, Verwalten, Strukturieren, Kombinieren, Präsentieren, Exportieren und Finden von Informationen.

Es gibt inzwischen eine Vielzahl solcher Systeme, die für unterschiedliche Zwecke konzipiert wurden und auch intern mit verschiedenen Konzepten und speziellen Datenformaten arbeiten.

In diesem Beitrag geht es darum, zu untersuchen, in welcher Weise sich die Extensible Markup Language (XML) für den Einsatz in Informationssystemen eignet. XML kann dabei sowohl als Format zur Speicherung von Dokumenten wie auch als Datenaustauschformat dienen.

Es kann hier keine Einführung in XML gegeben werden. Die Abbildung 1 soll nur einen Aspekt von XML-basierten Dokumenten veranschaulichen, der für das Folgende wichtig ist. XML-Dokumente entstehen als Zusammenspiel von mehreren Dateien.

Das Dokumentenmanagement von XML-Dokumenten steht also vor dem Problem, dass nicht nur eine Datei zu verwalten ist, sondern mehrere, die sich aufeinander beziehen. So gehören zu einer DTD in der Regel viele XML-Dateien. Andererseits kann es zu jeder XML-Datei mehrere Style-Dateien geben. Auch diese XSL- bzw. CSS-Dateien werden wiederum für verschiedene XML-Dateien benutzt.

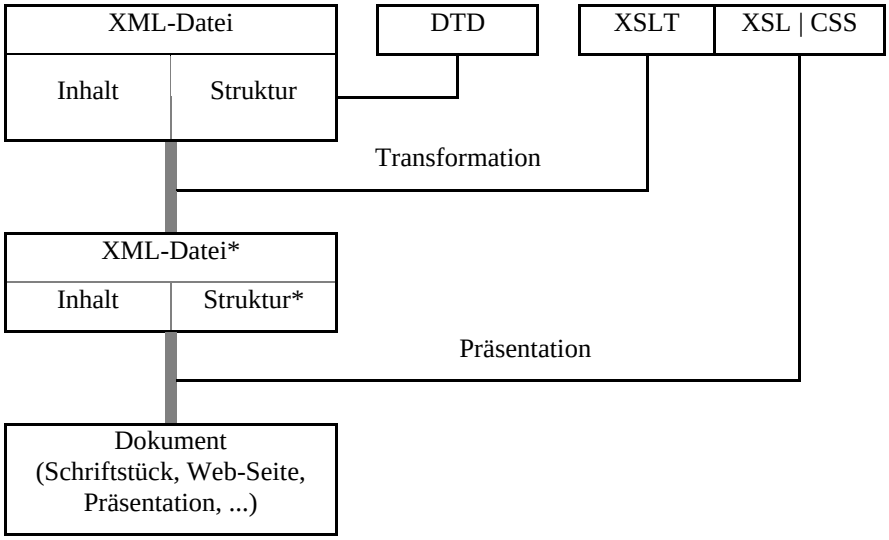


Abb. 1: XML-Dokumente

Obwohl dies die Dokumentenverwaltung etwas kompliziert, bietet XML andererseits eine Reihe von Vorteilen:

- XML ist ein herstellerunabhängiger Standard.
- XML bietet ein einheitliches Dokumentenformat, in das sich auch Multimedia-Objekte integrieren lassen.
- Medienbrüche lassen sich vermeiden, was die Publikationsmöglichkeiten verbessert.
- Das Format ist zukunftssicher und gut archivierbar.
- XML ermöglicht eine automatische Verarbeitung.
- XML bietet Möglichkeiten für eine intelligente und automatische Dokumentenindexierung.
- In XML-Dateien stehen implizite Metadaten zur Verfügung. Außerdem ermöglichen neue Standards, die Dokumente auch explizit mit Metadaten zu versehen.
- Auf der Basis der Auszeichnung der Dokumente mit Markups sind eine bequeme und zielgerichtete Suche und ein schneller Zugriff auf Dokumententeile möglich.
- XML-Dokumente sind gut zu vernetzen.

Die vielen genannten Vorteile stellen sich natürlich nicht im Selbstlauf ein, man muss in die Erstellung von XML-Dokumenten mehr Arbeit investieren als bei anderen. Die Festlegung der Dokumentenstruktur und das Erstellen der Markups kann nur zu einem geringen Teil automatisiert werden. Deshalb ist eine Schulung des Personals, das die Dokumente erstellt, unbedingt notwendig. Man muss sich zwingend an die vorgegebenen DTDs halten, die deshalb sehr gut durchdacht sein müssen. Derzeit stehen potentielle Nutzer vor dem Problem, dass es noch wenige Tools zur bequemen Erstellung von XML-Dokumenten gibt.

Will man XML als Dokumentenformat in Informationssystemen einsetzen, muss man sich u.a. über folgende Fragen Gedanken machen: Wie skaliert das System bei zunehmender Anzahl von Dokumenten und Nutzern? Wird der Mehraufwand für die Erstellung von XML-Dokumenten von den Informationsanbietern akzeptiert? Kann aus dem Einsatz von XML ein Mehrwert erzeugt werden - und mit welchem Aufwand? Wie ist die optimale Größe (Granularität) der Dokumente?

2. Wissenstransfer als Ziel von Informationssystemen

Ein wesentliches Ziel von Informationssystemen ist die Unterstützung des Wissenstransfers. Bevor wir dies näher betrachten, sollen zunächst einige Begriffe geklärt werden, die in der folgenden Diskussion auseinandergehalten werden müssen.

Informationen sind Nachrichten oder Berichte, die eine Bedeutung für den Adressaten haben. Ihre physische Ausprägung wird Daten genannt. Daten liegen im Computern in Form von Bits vor. Auf höherer Ebene sind auch XML-Dateien Daten.

Dokumente wiederum sind Daten, die vom Menschen rezipiert werden können. Sie sind mehr oder weniger strukturiert.

Wissen ist dagegen „verstandene“ Information. Es existiert nur im Gehirn des Menschen und besteht aus Fakten, Zusammenhängen, Strukturen, Klassifikationen und Erfahrungen. Wissen kann nur indirekt weitergegeben werden. Der Wissenstransfer ist die Weitergabe von Wissen in Form von Informationen. Es ist eine Transformation notwendig, die das Wissen einer Person in Dokumente (Buch, Vortrag, Video, ...) umwandelt. Diese Dokumente stellen für andere Personen Informationen dar, die i.allg. nur im zugehörigen Kontext verstanden werden können.

Um die Zusammenhänge zu veranschaulichen, kann man einen Wissenstransferzyklus [3] betrachten.

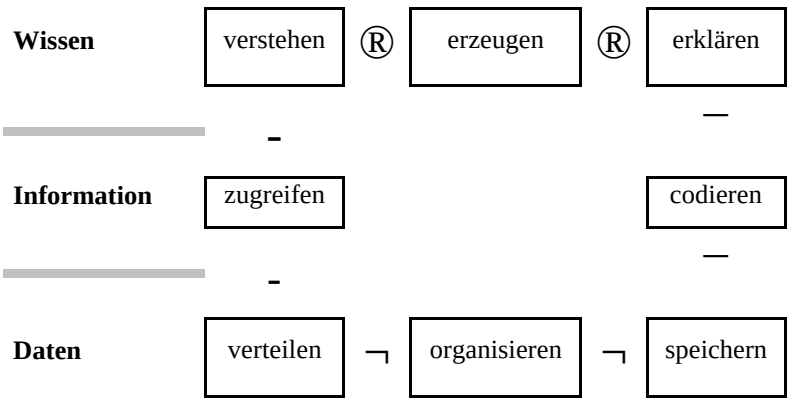


Abb. 2: Wissenstransferzyklus

In diesem Zyklus muss jedem Glied Beachtung geschenkt werden, sonst läuft der ganze Zyklus nicht rund. Häufige Schwachstellen sind die Phasen: erklären – codieren und organisieren – verteilen.

Wenn jetzt oft von Wissensmanagement die Rede ist, dann geht es darum, die oberste Ebene in dem betrachteten Kreislauf so zu organisieren, dass möglichst viel vorhandenes Wissen externalisiert, in Dokumenten codiert und damit für andere zugänglich gemacht wird. Aus oben genannten Gründen ist XML hierfür ein sehr gut geeignetes Datenformat.

Das Ziel des Knowledge Management (Wissensmanagement) ist, die Ressource *Wissen* besser zu nutzen. Das ist weniger ein technisches Problem als ein organisatorisches, psychologisches und firmenpolitisches. Das Wissensmanagement verlangt Akzeptanz bei den Angestellten, nutzerfreundliche Tools und eine darauf abgestimmte IT-Infrastruktur. Hier sind kaum Standardlösungen möglich, besonders dann nicht, wenn bestehende Systeme integriert werden sollen.

Die Abbildung 3 zeigt noch einmal den Wissenstransferzyklus. Dieser ist hier allerdings nun auf die Verwendung von XML-Dokumenten zugeschnitten.

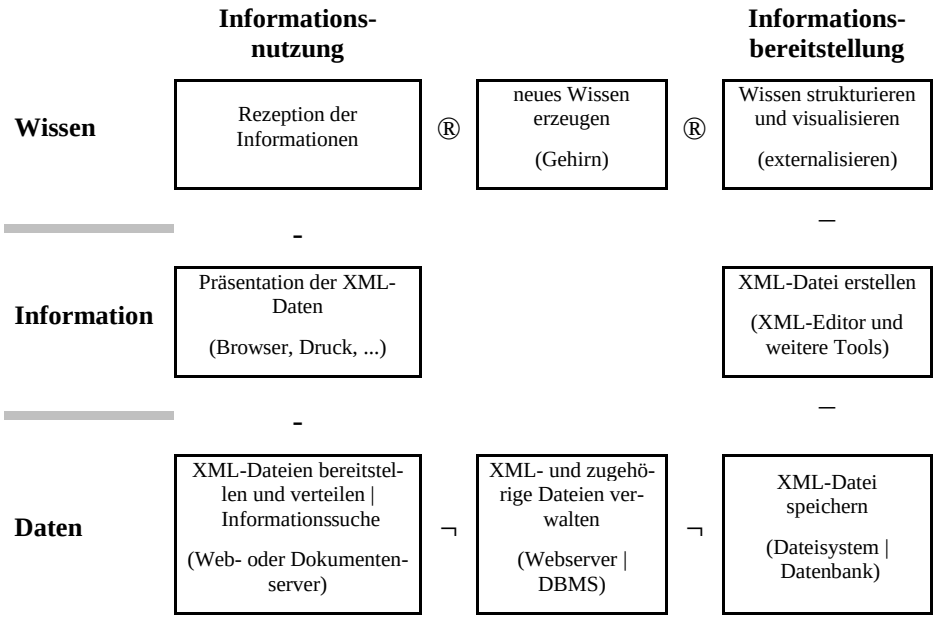


Abb. 3: Wissenstransferzyklus unter Benutzung von XML

Die Phasen in diesem Zyklus haben ganz unterschiedlichen Charakter. Es gibt eine Reihe besonders problematischer Phasen. Als erstes fällt sicher die Phase der Wissenserzeugung auf, die aber hier nicht betrachtet wird. Das ist kein Problem der Informatik. Auch die Aufgabe XML-Dateien zu erzeugen, ist technisch kein Problem, aber ein mentales. Hier kann Software mit einem guten Interface wichtige Hilfe leisten.

Im Zusammenhang mit dem Thema dieses Vortrages ist vor allem das Problem interessant, wie XML- und zugehörige Dateien verwaltet und verteilt werden.

Es wäre sicher übertriebener Ehrgeiz eine Lösung für das ganze Web entwickeln zu wollen. XML wird sich nur langsam durchsetzen, da es für unterschiedliche Anwendungen auch unterschiedliche Vorzüge besitzt. Ein Gebiet, wo XML große Chancen hat, ist das Online-Publishing von wissenschaftlichen Artikeln. Darauf soll im folgenden Abschnitt eingegangen werden.

3. Rolle von Metadaten beim Management von wissenschaftlichen Dokumenten

3.1. Wissenschaftlichen Dokumente

Bei wissenschaftlichen Dokumenten sind in Bezug auf die Verwaltung zwei Typen zu unterscheiden: statische und dynamische Dokumente.

Statische Dokumente repräsentieren einen abgeschlossenen Wissensstand (zum Fertigstellungszeitpunkt), sind zitierfähig und dürfen nicht mehr verändert werden. Derartige Dokumente sollten ausgiebig mit Metadaten beschrieben werden, da sie im allgemeinen für einen weltweiten ungehinderten Zugriff vorgesehen sind.

Dynamische Dokumente sind noch in Bearbeitung, unterliegen also ständigen Veränderungen. Diese Dokumente sollten für Gruppenarbeit geeignet sein und müssen deshalb durch eine Versionsverwaltung unterstützt werden. Hier muss auch der Zugriff geregelt werden, da dynamische Dokumente i.allg. nur für eine geschlossene Gruppe zugänglich sind. Bei diesen Dokumenten ist die Auszeichnung mit Metadaten noch nicht abgeschlossen. Die Dokumentenverwaltung muss hier viele Eigenschaften einer Groupware haben.

Das W3-Consortium hat durch die Entwicklung des *World Wide Web Distributed Authoring and Versioning* (WebDAV) [2], einer HTTP-Erweiterung für Teamarbeit im Web, die technische Grundlagen für die Bearbeitung von dynamischen Dokumenten im Web geschaffen.

Thesen:

1. Nicht die Verwaltung der Dokumente selbst ist das Problem, sondern Verwaltung der Metadaten (einschließlich Indizes, Beziehungen zwischen Dokumenten, Hyperlinks).
2. Die Volltextindizierung bzw. –suche ist kein optimales Mittel für die gezielte Suche in großen, heterogenen Datenbeständen.
3. Es ist sinnvoll, unterschiedliche Konzepte für die Erfassung wissenschaftlicher und „sonstiger“ Dokumente zu verfolgen. Auch die Behandlung statischer und dynamischer Dokumente muss auf unterschiedliche Weise erfolgen.
4. Für das Erstellen von wissenschaftlichen Dokumenten eignet sich XML in besonderer Weise (Gründe: herstellerunabhängiger und plattformübergreifender

Standard, gut archivierbar, Großteil der wissenschaftlichen Dokumente auf der Basis weniger DTDs anzufertigen).

5. Im Bereich wissenschaftlicher Dokumentenerstellung und -verwaltung sind Forderungen zur Schaffung von strukturierten und teilweise standardisierten Dokumente leichter durchzusetzen als auf anderen Gebieten. Dort ist das Bewusstsein um die Wichtigkeit seriöser Publikationen Allgemeingut. Wissenschaftler sind gewohnt, strukturiert zu denken und oft im Umgang mit formalen Sprache vertraut.
6. Die Indexierung von XML-Dateien würde durch DTDs mit einigen standardisierten Tags für Metadaten stark vereinfacht (z.B. <authors>, <title>, <keywords>, <year>, <institution>, ...). Die Metadaten könnten sich dabei z.B. an den Quasi-Standard Dublin Core anlehnen. [9]
7. Die Informationssuche und der wissenschaftliche Datenaustausch könnten mit Hilfe von XML beschleunigt und effektiviert werden.

3.2. Indexierung von Dokumenten

Um Dokumente mit gewünschten Informationen schnell zu finden, müssen Angaben über diese Dokumente (Metadaten) an einer zentralen Stelle gesammelt werden. Dort werden Indizes aufgebaut, die einen schnelle Zugriff auf die ursprünglichen Dokumente ermöglichen. Unter Index wird hier eine Datei verstanden, die Informationen (Metadaten) über viele Dokumente in einer strukturierten Form enthält.

Folgende Phasen werden beim Aufbau von Indizes durchlaufen:

1. Metadaten aus den Dokumenten ermitteln (auslesen, zuordnen)
2. *optional:* Metadaten separat speichern (z.B. als XML- oder RDF-Datei)
3. Aufbau von Indizes aus den Metadaten
4. Verwaltung und Aktualisierung der Indizes

Die Abbildung 4 veranschaulicht das für den Fall von XML-Dateien. Die Indexierung läuft hier in zwei Schritten ab. Im ersten werden die Metadaten aus den Dokumenten extrahiert und in Metadaten-Dateien gespeichert, während im zweiten diese Dateien ausgewertet und Indizes erstellt werden.

Die Metadaten-Dateien sollten auf der Basis einer einheitlichen DTD wiederum im XML-Format erstellt werden. Dabei können die Metadaten mehrerer Dokumente in einer Datei gespeichert werden. Das bietet sich besonders bei XML-Dateien an, die auf der gleichen DTD beruhen.

Im zweiten Schritt dienen die Metadaten-Dateien zum Aufbau von Indizes, die Informationen über alle Dokumente der betrachteten Domäne enthalten.

Diese zweistufige Indizierung, die auch beim Harvest-System (lokaler *Gatherer* und zentraler *Broker*) benutzt wird [1], hat m.E. verschiedene Vorteile gegenüber einem einstufigen Vorgehen:

1. Die Erstellung der Metadaten-Dateien kann verteilt geschehen. Jeder Server erstellt selbstständig diese Dateien und übermittelt sie an den Index-Server.

2. XML-Dateien eignen sich nicht nur gut zur Speicherung von Metadaten, sondern sind auch gut für die Übertragung über Datennetze geeignet.
3. Der Index-Server muss lediglich die Metadaten-Dateien auswerten, die in einem einheitlichen Format vorliegenden sollten. Er kann basierend auf diesen Daten verschiedene Indizes aufbauen.
4. Es können auch Dokumente erfasst werden, die nicht im XML-Format vorliegen, indem spezielle Tools aus diesen die Metadaten extrahieren und im XML-Format ablegen.
5. Vorhandene Indizes lassen sich integrieren, indem diese ins XML-Format exportiert werden.
6. Die Metadaten können auch manuell erstellt oder überarbeitet werden.

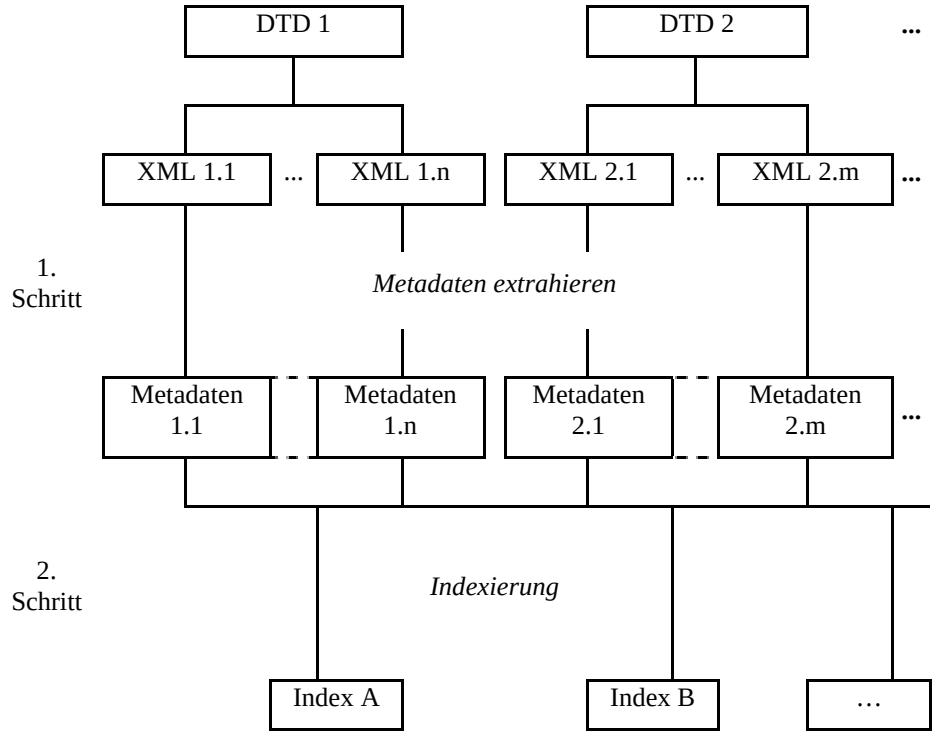


Abb. 4: Zweistufige Architektur zur Erstellung von Indizes

Die folgende DTD könnte XML-Dateien zur Speicherung von Metadaten zugrunde liegen:

```
<!--Sammlung von Metadaten ueber Dokumente-->
<!ELEMENT metadata (index, ducument*)>
<!--Hauptabschnitt "index"-->
```

```
<!ELEMENT index (date_change, index_id+, generator+,
description+)>
<!ELEMENT date_change (#PCDATA)>
<!ELEMENT index_id (#PCDATA)>
<!ELEMENT generator (#PCDATA)>
<!ELEMENT description (#PCDATA)>
<!--Hauptabschnitt "ducument"-->
<!ELEMENT ducument (id, keywords, authors+, content)>
<!--Abschnitt "id"-->
<!ELEMENT id ((uri | isbn | special_id),
date_ducument+, date_index)>
<!ELEMENT uri (#PCDATA)>
<!ELEMENT isbn (#PCDATA)>
<!ELEMENT special_id (#PCDATA)>
<!ELEMENT date_ducument (#PCDATA)>
<!ELEMENT date_index (#PCDATA)>
<!--Abschnitt "keywords"-->
<!ELEMENT keywords (keyword)*>
<!ELEMENT keyword (#PCDATA)>
<!--Abschnitt "authors"-->
<!ELEMENT authors (firstname*, surname, institution+,
address+)*>
<!ELEMENT firstname (#PCDATA)>
<!ATTLIST firstname Typ (full | abbreviate) #IMPLIED>
<!ELEMENT surname (#PCDATA)>
<!ELEMENT institution (#PCDATA)>
<!ELEMENT address (#PCDATA)>
<!--Abschnitt "content"-->
<!ELEMENT content (title, subtitle+, abstract*,
structure+)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT subtitle (#PCDATA)>
<!ELEMENT abstract (#PCDATA)>
<!ATTLIST abstract language (de | en) #REQUIRED>
<!ELEMENT structure (chapter)*>
<!ELEMENT chapter (#PCDATA)>
```

Jede auf dieser Document Type Definition beruhende XML-Datei besteht aus zwei Hauptteilen. Der erste Teil, von <index>-Tags umschlossen beschreibt die Datei insgesamt, der zweite enthält die Metadaten beliebig vieler Dokumente, jeweils von <document> geschachtelt. Die Beschreibung der Dokumente hat wiederum vier Abschnitte, die das Dokument eindeutig identifizieren, Stichworte angeben, den oder die Autoren benennen und den Inhalt charakterisieren.

Diese DTD lässt sich noch in verschiedener Richtung ausbauen, um weitere Aspekte abzudecken. Dabei sollten auch Standards wie Dublin Core [9] oder das Resource Description Framework (RDF) berücksichtigt werden [7].

Die Sammlung und Verarbeitung von Metadaten ist allerdings nur ein Aspekt bei der Strukturierung von Informationssystemen.

4. Infrastruktur eines verteilten Informationssystemen

Bei der folgenden Betrachtung geht es nicht um geschlossene Informationssysteme, die alle Komponenten enthalten, sondern um Systeme, die verteilt sind. Das bedeutet insbesondere, dass die sich Dokumente und Indizes nicht alle auf einem Server befinden, sondern auf viele Maschinen verteilt sind. Die Skala der möglichen Systeme reicht dabei von einem einzelnen Nutzer, der privat Information organisieren möchte, bis zum World Wide Web. Natürlich lassen sich diese unterschiedlichen Systeme nicht mit dem gleichen Ansatz bewältigen. Hier soll ein Informationssystem betrachtet werden, das zum Beispiel in einem Intranet zum Einsatz kommen könnte.

Ein Informationssystem enthält i.allg. folgende Hauptkomponenten:

Autorenunterstützung	**Verwaltung**	**Nutzerunterstützung**
Editor bzw. Entwicklungsumgebung	Datenspeicherung (einschl. Backup und Versionsverwaltung)	Suche
Import-Schnittstelle	Indexierung	Datenstrukturierung
Metadatenerfassung	Archivierung, Löschen	Präsentation, Export

Bei dem zu betrachtenden System liegen die Daten auf Dokumenten- oder Web-Servern, was zur Folge hat, dass die Autorenunterstützung und die Datenspeicherung nicht der Kontrolle des eigentlichen Informationssystems unterliegen.

Die Struktur eines relativ einfachen, verteilten Informationssystems könnte wie in Abbildung 5 aussehen. Das Informationssystem besteht dabei im Kern aus dem Informationsassistenten und den kommunizierenden Index-Servern.

Der Nutzer wendet sich mit seinen Wünschen an den Informationsassistenten. Dieser hat nun die Aufgabe, die gewünschte Information zu suchen und dem Nutzer zu präsentieren. Bei seiner Suche wendet sich der Assistent an einen oder mehrere Index-Server, die über Angaben verfügen, welche geeigneten Dokumente existieren und wo sie sich befinden. Der Nutzer bekommt i.allg. zunächst Metadaten angezeigt, woraufhin er seine Anfrage ändern oder eine Auswahl treffen kann.

Im einfachsten Fall werden die gefundenen Dokumente bzw. dessen Positionen an den Nutzer weitergeleitet. Diese Funktionalität bietet allerdings auch bereits eine Suchmaschine.

Um ein Mehrwert-Informationssystem zu realisieren, werden an den Informationsassistenten höhere Anforderungen gestellt. Er nimmt die Dokumente entgegen, filtert, kombiniert und präsentiert die Informationen nach den Wünschen des Nutzers. XML bietet dazu ein geeignetes Datenformat. Es bleibt allerdings noch genauer zu untersuchen, welche Anforderungen an die XML-Dokumente gestellt werden müssen, damit eine weitgehend automatische Verarbeitung möglich ist.

Der Datenaustausch zwischen den Komponenten des obigen Systems kann in vertikaler Richtung mit Hilfe von HTTP erfolgen. Für den Austausch zwischen den Index-Servern, die eine verteilte Datenbasis bilden, muss ein eigenes Protokoll entwickelt werden, dessen Anforderungen noch genauer zu spezifizieren sind.

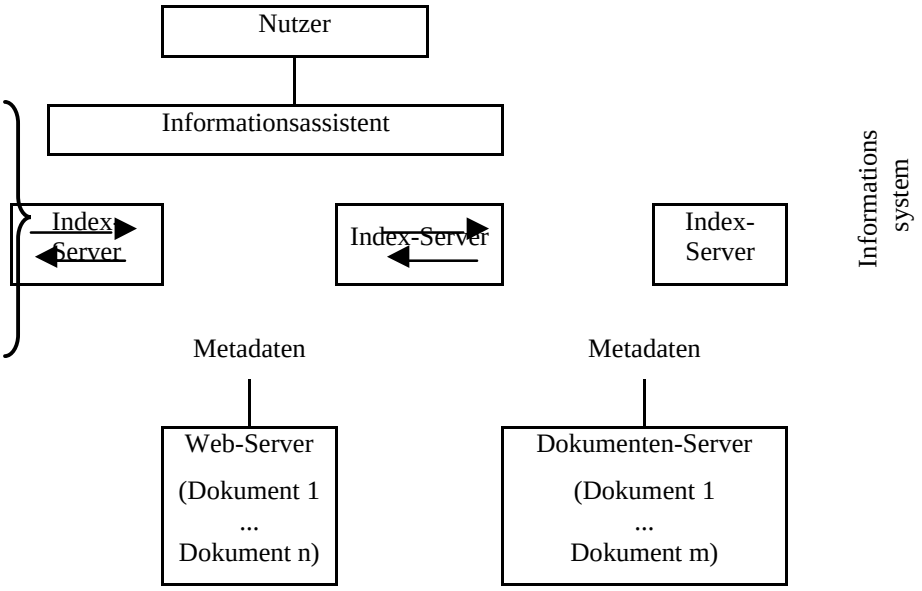


Abb. 5: Infrastruktur eines verteilten Informationssystems

5. Zusammenfassung

XML hat die Potenzen, das Dokumentenmanagement in verteilten Informationssystemen zu effektivieren und dem Nutzer solcher Systeme einen deutlich verbesserten Service zu bieten. Ein Haupthindernis für die schnelle Durchsetzung dürfte jedoch sein, dass das Erstellen von XML-Dokumenten mit einem erhöhten Aufwand verbunden ist. Die exakte Strukturierung der Dokumente stellt hohe Anforderungen an die Autoren. Deshalb scheinen sich im Umfeld des wissenschaftlichen Arbeitens und Publizierens die besten Voraussetzungen für die Durchsetzung von XML als Dokumentenformat zu bieten. Bewährt sich die Extensible Markup Language für Informationssysteme in diesem Bereich, dann dürften sich auch schnell weitere Anwendungsfelder bieten.

6. Literatur

[1] Harvest Project: http://harvest.transarc.com/

[2] IETF WebDAV Working Group: http://www.ics.uci.edu/pub/ietf/webdav

[3] F. Kappe: Aufbau und Nutzung von Wissenspotentialen in verteilten Organisationen, Vortrag auf der Tagung „Online '99", Düsseldorf, 02.02.1999

[4] Ch. Koch: Konzeption und prototypische Implementierung von Softwarekomponenten zur strukturierten Ablage von Informationsobjekten in einem Repository, Diplomarbeit an der TU Ilmenau 1997

[5] M. A. Lopéz Vargas: Informationsfilterung in einem verteilten Information Retrieval System, Vortrag auf dem 5. Workshop "Multimediale Informations- und Kommunikationssysteme", Erfurt 30.09.1999

[6] P. Reimann, K. Müller, Ph. Starkloff: Kognitiv kompatibel? – Wissensmanagement zwischen Technik und Psyche, c't 4/2000 S.274ff.

[7] E. Schön: Das Resource Description Framework (RDF) - ein neuer Weg zur Verwaltung von Metadaten im Netz, Vortrag auf dem 4.Workshop „Multimediale Informations- und Kommunikationssysteme", Ilmenau 09.10.1998

[8] G. Versteegen: Was ich nicht weiß ..., iX 7/2000, S.100ff.

[9] S. Weibel: The State of the Dublin Core Metadata Initiative, D-Lib Magazine, April 1999, Volume 5 Number 4, http://www.dlib.org/dlib/april99/04weibel.html