Der ELSA-Prototyp aus informationstechnischer Sicht

Heinz-Dirk Luckhardt

Ausgangslage

State of the Art

(Elektronisches)Rohmaterial

Motivation für die Informationswissenschaft

Informationstechnische Alternativen

Geplante Systemeigenschaften

Die ELSA-Website

****

Ausgangslage

State of the Art, eigene Vorarbeiten

- Das WWW als Informations- und Kommunikationsmedium

In den knapp 10 Jahren seines Bestehens hat sich das World Wide Web zu einem universell verfügbaren, unüberschaubar großen Informations- und Kommunikationssystem entwickelt. Diese Entwicklung schließt ständige große Veränderungen in der Nutzung und bei den Nutzern, in Aussehen, Inhalten, technischen Grundlagen usw. ein, praktisch in Bezug auf alle Aspekte und Komponenten.

- Stand der Entwicklung des WWW im Jahre 2000: Schlagwort „Content!“

Waren es noch zu Beginn in erster Linie einzelne engagierte Universitätsangehörige, die praktisch ohne Auftrag, fasziniert von den sich abzeichnenden Möglichkeiten des neuen Mediums sie selbst interessierende Inhalte ins WWW brachten, so sind es heute Wirtschaftsunternehmen, die das WWW „übernommen" haben. Das Engagement Einzelner kann sicher nicht Aufbau und Bestand eines universellen, sich ständig fortentwickelnden Informationssystems garantieren. Und so sind neue Vorgaben, Modelle, Methoden, Kooperationsstrukturen und Verfahrensweisen nötig, um diesem Zweig des Elektronischen Publizierens auf Dauer zum Erfolg zu verhelfen. Im Bereich der Bibliotheken gibt es hier schon richtungsweisende Entwicklungen, z.B. hier an der Saarländischen Universitäts- und Landesbibliothek das Saarbrücker Online-Volltextarchiv SOVA .

- Einstieg in das Gebiet der elektronischen Online-Archive

Die Fachrichtung Informationswissenschaft hat in den letzten drei Jahren einige Vorarbeiten geleistet, deren Ergebnisse im Projekt ELSA genutzt werden können.

Seminar 1999/2000: *Terminologie-Datenbank T-REX (Terminosaurus Rex)*

Das Besondere an T-REX ist der datenbankgestützte Ansatz, der es ermöglicht, große Mengen strukturierter Daten zu speichern und auf verschiedenen Wegen verfügbar zu machen. Hier handelt es sich um einige hundert informationswissenschaftliche Fachbegriffe mit Definitionen, Literaturverweisen, einem Begriffsnetz und interessanten Beiträgen im WWW.

zwei Seminare 1998/99 zum Werk von Alfred Gulden: HTML-Projekte, Werkschau als Datenbank

Hauptziel der Seminare mit und über Alfred Gulden war es, die Möglichkeiten des WWW für die Darstellung verschiedener künstlerischen Ausdrucksformen (Roman, Lyrik, Film, Kurzgeschichte ...) auszuloten. Herausgekommen ist eine Sammlung von Darstellungen, die über die Gulden-Website im WWW zugänglich ist. Das Besondere – im Hinblick auf die Verwertung der Ergebnisse in ELSA – war die individuelle Gestaltung von Webangeboten mihilfe der HTML-Technologie und besonderer Gestaltungstechniken (Javascript, Flash, Shockwave), aber auch ein datenbankgestütztes Verzeichnis der Werke Guldens.

Seminar 2000/01 über Friedrich Nietzsche: Annäherung an die XML-Welt

Schließlich möchte ich noch das Nietzsche-Seminar aus dem Wintersemester 2000/2001 nennen, das sich die Erstellung eines Online-Angebots über Friedrich Nietzsche mit Hilfe der XML-Technologie zur Aufgabe gemacht hat. Sie sehen den Ausschnitt einer Webseite, die den Zugang zu häufig gestellten Fragen (und Antworten dazu) über eine Schlagwortliste bietet. Die Anwendung der XML-Technologie bietet sich hier an, weil damit die inhaltliche Strukturierung und die detaillierte Verknüpfung von „Informationsobjekten" (Briefen, Texten, Personenbeschreibungen, Frage und Antwort etc.) in idealer Weise möglich ist.

- Verfügbare Technologien: HTML, XML, Datenbank-Ansatz

Die drei genannten Beispielprojekte repräsentieren also die wichtigsten im WWW eingesetzten Technologien HTML, XML und Datenbank, die im Projekt ELSA auf ihre Tauglichkeit im Bereich „elektronisches Archiv" überprüft werden.

(Elektronisches) Rohmaterial

Der Output des ersten Arbeitsgangs, die bibliographische Erfassung und das Scannen der Archivobjekte, stellt das Rohmaterial für das Online-Archiv dar. Um welche Objekte es sich handelt, wird im Beitrag von Hermann Gaetje erläutert.

- Große Mengen gleich strukturierter Objektbeschreibungen

Es sind zum einen die Objektbeschreibungen, die nach vorgegebenen Mustern bzw. Formularen erfaßt werden.

- Strukturierte, verknüpfbare Texte (z.B. Briefe)

Zum anderen sind dies die durch Scannen, OCR (= optical character recognition) bzw. von Hand erfassten Texte (Briefe, Sekundärliteratur) und die von anderen Medien in digitale Objekte umgewandelten Archivobjekte (Bilder, Fotos, Tondateien, Videos).

Motive der Informationswissenschaft

Für die Informationswissenschaft ist das Projekt ELSA aus verschiedenen Gründen von großem Interesse.

- Analyse, Planung, Entwicklung unterschiedlicher Informationssysteme

Die Arbeit von Informationswissenschaftlern kann man sich als die von Informationsarchitekten vorstellen, deren Aufgabe die Planung und Entwicklung der unterschiedlichsten Informationssysteme ist. Wie der Bau eines Hauses ist auch die Entwicklung eines Informationssystems eine so komplexe Aufgabe und erfordert so viele unterschiedliche Kompetenzen, dass die wichtigste Aufgabe vor allem in der Koordinierung der Arbeiten liegt. Jedes neue Informationssystem und jeder neue Typus „Informationssystem“ bietet neue Herausforderungen.

- Erprobung neuer Informationstechnologien

Im Bereich Forschung und Entwicklung ist vor allem die Erprobung neuer Informationstechnologien von Interesse. Die drei geschilderten Technologien sind alle noch so neu und entwickeln sich ständig so rasch fort, dass ihre Anwendung auf neue Aufgaben eine Daueraufgabe der Informationswissenschaft ist.

- Benutzungs- und Bedarfsanalyse: redundante Zugangsmöglichkeiten

Im Mittelpunkt des informationswissenschaftlichen Interesses hat immer der Mensch als Nutzer von Informationssystemen gestanden. Das bedeutet, dass ein besonderes Augenmerk auf der Benutzbarkeit der zu entwickelnden Systeme liegt. „Redundante“ Zugangsmöglichkeiten heisst hier, dass die Systembenutzer unterschiedliche, also individuelle Wege zur Information verfolgen können, im Sinne von „Viele Wege führen nach Rom“.

- Information Retrieval: Suchen und Finden in großen Datenmengen

Der ELSA-Prototyp aus informationstechnischer Sicht

Heinz-Dirk Luckhardt

Traditionell liegt ein Hauptaugenmerk der informationswissenschaftlichen Arbeit auf Verfahren der Informationssuche in großen Datensammlungen. Es hängt von den Voraussetzungen der Informationssuchenden ab, welche Verfahren eingesetzt werden können:

- Suche in geordneten Listen (Bestandsübersicht)

Wenn ein Informationssuchender seine Frage nicht präzise formulieren kann oder sich einen Überblick über das vorhandene Material verschaffen möchte, bietet sich die Suche in geordneten Listen an. Diese enthalten die Beschreibungen der Archivobjekte, nach Typ geordnet, und verweisen auf die mit ihnen verbundenen elektronisch erfassten Texte, Bilder, Tonaufnahmen etc.>

- Suche in einem XML-Netz

Die XML-Technologie bietet optimale Voraussetzungen für die Verbindung strukturierter Texte zu einem Netz. Der Vorteil für die Nutzung liegt darin, dass die Suche nach Informationen auf bestimmte Suchräume eingeengt werden und gleichzeitig die Nachbarschaft von Dokumenten dargestellt und ausgenutzt werden kann, z.B. die Relation zwischen Briefen (Vorgänger, Nachfolger, erster Brief eines ganzen Briefwechsels ...).

- Stichwortsuche

Wenn ein Informationssuchender einen präzisen Suchbegriff nennen kann, bietet sich die Stichwortsuche an, die den Suchbegriff in bestimmten Textfeldern der Datenbank sucht und die Dokumente nennt, die den Begriff enthalten (vgl. die Beiträge zu Information Retrieval und zur Indexierung im Online-Handbuch der Informationswissenschaft). Diese Fragestellung ist insofern außerordentlich komplex, als die Suchfunktion gleichzeitig sehr einfach zu bedienen und sehr leistungsstark sein soll. Auf der einen Seite soll sie die Möglichkeit bieten, die Suchanfrage detailliert zu spezifizieren, also möglichst genau das inhaltliche Profile der gesuchten Dokumente anzugeben, auf der anderen Seite muss die Suchfunktion auch für Laien verständlich sein, was in der Regel nur die Eingabe einfacher Suchbegriffe erlaubt. Diesem Aspekt wird im Projekt noch durch eine Aufspaltung der Suchfunktion in eine einfache und eine komplexe Funktion Rechnung getragen werden.

Eine andere wichtige Frage ist die, inwiefern Stichwörter geeignet sind, den Inhalt eines Dokument zu beschreiben: soll man nicht eher mit Schlagwörtern - also normierten Bezeichnungen - arbeiten, anstatt nur mit Wörtern aus dem Text (Stichwörtern), mit allen bekannten Problemen durch Synonymie und Homonymie? In dieser ersten Projektphase lässt sich sich diese Frage nur damit beantworten, dass eine Beschreibung der Objekte mit (vorgegebenen) Schlagwörtern weit über die Möglichkeiten des Projekts hinausgehen.

- Thematische Verknüpfung (wiederkehrende Motive):

Neben den traditionellen Retrievalmöglichkeiten ist die Idee eines thematischen Netzes entstanden, das eines der Werkzeuge sein könnte, um das Archivmaterial zu durchsuchen. Dies gründet darauf, dass sich bei dem Autor Gulden verschiedene wiederkehrende Motive finden (Grenze, Welt und Winkel, Identität und Distanz ...). Für das Thema „Grenze“ ist der Versuch gemacht worden, relevante Dokumente zu einem Netz zu verbinden und über eine zunächst noch primitive Schnittstelle zugänglich zu machen.

Informationstechnische Alternativen

Die informationstechnischen Alternativen haben ihre Vor- und Nachteile, die im Folgenden kurz miteinander verglichen werden sollen.

Systemunabhängigkeit

Nur XML ist systemunabhängig. Es ist ein offenes Format, das nicht von der Systemplattform abhängt. Bei den Datenbanken muss man sich für einen Typ entscheiden (obwohl es hier natürlich Schnittstellen zwischen verschiedenen Typen gibt). HTML birgt in dieser Hinsicht einige Probleme, da nicht alle Elemente system- d.h. browserunabhängig sind.

Erweiterbarkeit

XML ist auf Erweiterbarkeit angelegt, HTML kann nicht individuell erweitert werden. Datenbanken sind in ihrer Struktur – wenn sie einmal angelegt sind – mit Einschränkungen erweiterbar.

Strukturierbarkeit

Informationsobjekte lassen sich mit XML und in Datenbanken inhaltlich gut strukturieren. HTML bietet nur die Möglichkeit der optischen Strukturierung. Dieser Aspekt wirkt sich besonders auf die gezielte inhaltliche Suche in strukturierten Objekten aus, d.h. XML und Datenbanken haben hier große Vorteile gegenüber HTML.

Aufwand

Mit HTML sind schnell WWW-Angebote zu erstellen, die aber dann einen relativ hohen Aufwand an Wartung erfordern. XML erfordert einen sehr hohen Anfangsaufwand (Analyse der Dokumentenmenge, Entwurf der DTDs, Erstellung / Generierung der XML-Dateien, Programmierung der Formatierungsfunktionen). Änderungen lassen sich dann relativ rasch und systematisch durchführen. Datenbanken erfordern informatisches Know-How, d.h. Entwickler müssen in Informatik geschult sein.

Ausgabe auf unterschiedliche Medien

XML-Dokumente lassen sich in die unterschiedlichsten Ausgabe- und Austauschmedien überführen, auch Datenbanken bieten hier einige Möglichkeiten. HTML-Dokumente sind in dieser Hinsicht äußerst problematisch.

Verknüpfung von Informationsobjekten

Die Verknüpfung von Dokumenten oder Dokumentteilen ist mit allen drei Alternativen leicht möglich, wobei HTML die flexibelsten und XML bzw. Datenbanken systematischere Verknüpfungsmöglichkeiten bieten.

Nutzung der Vorteile der unterschiedlichen Technologien

Ausgehend von den vorstehenden Überlegungen wurden in Abhängigkeit vom Objekttyp die Entwurfsentscheidungen für das elektronische Archiv getroffen.

DB: Datenbanken sind von Vorteil, wenn es um die Verwaltung großer Mengen gleichartiger Daten geht. Damit ist

- die automatische Konversion von Daten mit einfachen Programmen,
- einfache Zugriffsfunktionen über automatische erzeugte Linklisten,
- die einfache Suche in definierten Datenfeldern

leicht möglich.

XML: Der Einsatz von XML bietet sich in erster Linie bei grossen Datenmengen an, die inhaltlich strukturiert werden müssen und die ständig erweitert und gepflegt werden sollen. Mit XML sind die flexiblere Strukturierung und eine leichtere und komplexere Verknüpfung von Elementen als in Datenbanken möglich.

HTML: Mit HTML ist die einfache Darstellung weniger strukturierter, adhoc zu verknüpfender Sachverhalte möglich.

Geplante Systemeigenschaften

Das elektronische Archiv soll nach der ersten Projektphase den folgenden Anforderungen genügen:

einfache Datenerfassung

Die bisherigen Arbeiten habe gezeigt, dass die gewählten Verfahrensweisen eine einfache Erfassung der unterschiedlichen Daten ermöglichen.

einfache Überführung der erfassten Daten in andere Formate

Am Ende der ersten Phase werden Prozeduren entwickelt sein, mit denen in Zukunft der Ausbau der Datenbanken problemlos möglich sein wird.

einfacher Zugang zu allen Funktionen über WWW-Oberfläche

Der Zugang zum Archiv über eine Website steht in einer ersten Version zur Verfügung, die entsprechend den Möglichkeiten des WWW jederzeit angepasst und ausgebaut werden kann.

Erweiterbarkeit um neue Objekt-/Dokumenttypen, Inhalte, Funktionen

Da sich die Verfahren zur Erfassung und Umsetzung für die bisher berücksichtigten Objekttypen bewährt haben, lassen sich auf ihrer Grundlage bis zum Ende dieser Projektphase Standards für die Einbettung neuer Objekttypen erstellen.

Portierbarkeit auf andere Betriebssysteme

Das ELSA-System, das derzeit unter LINUX läuft, ist jederzeit auf andere Plattformen portierbar, die die Einrichtung von Webservern mit PHP-Laufzeitumgebung erlauben.

Die ELSA-Website

Die Eingangsseite: ein Blick ins Archiv

Das ELSA-Archiv empfängt die Besucher auf der Begrüßungsseite mit einer Photographie aus dem Archiv, die Alfred Gulden bei den Vorarbeiten zu seinem Roman „Die Leidinger Hochzeit“ aufgenommen hat. Die Entscheidung für dieses Photo hat das Design der gesamten Website beeinflusst, da aus ihm die identitätsstiftende Farbgebung abgeleitet wurde, d.h. alle Farben entstammen dem Bild. Diese Vorgehensweise, d.h. die besondere Hervorhebung eines einzelnen Archivobjekts bzw. eines besonderen Motivs, ist sicher diskussionswürdig, lässt sich aber auch bei anderen Archiven beobachten.

Die Systemfunktionen

Nach wenigen Sekunden wird man auf die Inhaltsseite geleitet, von der aus alle Systemfunktionen erreicht werden. Neben Informationen über das Projekt, den Vorlass Alfred Guldens und den technischen Hintergrund ist hier auch die eigentliche Funktion der Website erreichbar, nämlich die Suche im elektronischen Archiv.

Die Suchfunktionen

Von den vier Suchfunktionen wurden die ersten beiden datenbankgestützt, die dritte mit XML und die vierte mit HTML realisiert.

Suche in Listen von Objekten (Bestandsübersicht)

Die Objektlisten, ihre Inhaltsverzeichnisse und die Verweise auf digitalisierte Objekte wie Texte oder Bilder werden aus der Datebank dynamisch erzeugt.

Suche nach bestimmten Wörtern in allen Objektbeschreibungen

Bei der Stichwortsuche kann über ein Suchformular ein zu suchendes Wort eingegeben werden. Die Suche erfolgt nicht im Volltext, sondern in den Kurzbeschreibungen der Objekte.

Suche nach strukturierten/vernetzten Objekten (Briefen)

Für einen Ausschnitt der Briefsammlung wurde eine XML-Anwendung erstellt, die exemplarisch die Möglichkeiten von XML zeigt: die inhaltliche Strukturierung von Objekten und ihre Verknüpfbarkeit mit anderen Objekten/Objektbeschreibungen.

Stöbern in einem thematischen Netz

Die Idee eines thematischen Netzes wurde für das Thema „Grenze“ anhand einiger Texte bzw. Archivmaterialien über dieses Thema realisiert.