Aufgabenblatt 2

1. Begriffe

Biogramm Ein kurzer Biographischer Abriss, in unserem Falle vor allem Informationen über die Universitätslaufbahn.

Buchdruckfunktion Möglichkeit, statt der direkten Ausgabe am Bildschirm eine Datei mit den ausgewählten Daten erstellen zu lassen. Diese soll unter möglichst geringem Weiterverarbeitungsaufwand das Publizieren von auswählbaren Teilen (prinzipiell natürlich auch aller Daten) der Datensammlung in Buchform ermöglichen.

LaTeX Weiterentwicklung des TeX-Textsatzsystems (*siehe Konzepte*).

OntoWiki ist eine webgestützte Software zum Erzeugen, Manipulieren und Durchsuchen von Ontologien bzw. den dahinter stehenden Daten.

Professorenkatalog (allgemein) Datensammlung von Hochschullehrern (nicht nur ordentlicher Professoren, sondern auch PDs, außerordentlicher Prof., usw.)

Professorenkatalog (dieses Projekt) Frontend zum Durchsuchen, Anzeigen und Ausgeben der Datenbank des Professorenkatalogs der Universität Leipzig.

Recherche bezeichnet das gezielte Suchen von Daten im Professorenkatalog. Dabei wird jedoch nur ein Teil der vorhandenen Daten nach außen durchsuchbar und sichtbar sein.

2. Konzepte

Die Erfolge in Forschung und Entwicklung in der Digitaltechnik der letzten Jahrzehnte ermäglichen es, immer mehr und immer komplexere Daten zu Speichern und zu Verbreiten. Relevanten Daten zur Verarbeitung und Anwendung in den wachsenden Datenbeständen aufzufinden wird folglich immer schwieriger. Nun werden die syntaktischen Möglichkeiten zur Datenstrukturierung (d.h. Dateien in Dateisystemen und Daten in Datenbanksystemen mit bestimmten Formaten, oder Sprachen) jedoch unzulänglich, da durch sie die Bestimmung der Relevanz der Daten nicht gewährleistet werden kann. Folglich wird die logische Erweiterung dazu angestrebt, nämlich semantische, also kontextbezogene, Beziehungen computergestützt verarbeiten zu können (im Rahmen einer WWW-Anwendung wird dann vom *Semantic Web* gesprochen). Dazu wird ein Modell des Bereiches, aus dem die Daten stammen, erstellt, in dem sich diese Beziehungen widerspiegeln. Das Modell muss, um von Computern verarbeitet werden zu können, formalisiert sein. Das Modell ist weiterhin allen Instanzen, die auf den Daten operieren bekannt. Solche Modelle werden *Ontologien* genannt. Ontologien können weiterhin beliebig um neue Aspekte erweitert werden, was sie besonders interessant für die konkrete Anwendung im Professorenkatalog macht. Wird vom Kunden die Erfassung weiterer Aspekte gewünscht, so kann dies leicht durch die Anpassung der zugrundeliegenden Ontologie erreicht werden.

Ein Beispiel für eine konkrete Ontologie:
Dublin Core (DC) ist ein Metadaten-Schema zur Beschreibung von Dokumenten und anderen Objekten im Internet. Autoren von Webressourcen sollen durch dieses Metadatenschema in die Lage versetzt werden, ihre Ressourcen so zu beschreiben, dass sie etwa von stichwortbasierten Suchmaschinen gefunden werden können. Es ist das Ergebnis einer jahrelangen Diskussion zwischen Informatikern, Wissenschaftlern und Bibliothekaren. Das Dublin Core Metadata Set besteht aus 15 Elementen zur Beschreibung von elektronischen Dokumenten. Diese Elemente können durch Subklassen qualifiziert werden. Durch die simple Form der Beschreibung ist DC sehr einfach und effektiv einsetzbar.
Um einem Dokument Metadaten hinzuzufügen wird ihm zunächst eine eindeutige ID zugewiesen, beispielsweise eine URL oder eine ISBN, d.h. eine Identifizierung durch einen Katalog. Weiter gibt es Felder für technische Daten (Format, Typ, Sprache), Beschreibung des Inhalts (Titel, Thema, Eingrenzung des Themas, Kurzzusammenfassung), Personen und Rechte (Verfasser, Herausgeber, Rechte), Vernetzung (inhaltliche Quellen, Dokument mit Bezug zum Aktuellen; hier sind vor allem IDs anderer mit Metadaten versehener Dokumente gemeint) und ein Datum zum Lebenszyklus. Die Darstellung ist beispielsweise im Format RDF möglich, jedoch nicht darauf beschr"nkt („interoperability“), ebenso kann DC direkt in HTML-Dokumente eingebettet werden. Damit wird im Optimalfall ein Beziehungsgeflecht zwischen den über das WWW zugänglichen Dokumenten aufgestellt, das alle Seiten über ihre Eigenschaften an den Stellen größerer Relevanz miteinander verbindet.

Zur Formulierung von Ontologien gibt es verschiedene Ansätze und Sprachen. In unserem Projekt wird dafür die vom W3C erarbeitete Sprache *Resource Description Framework* (RDF) sowie deren Erweiterungen RDF-Schema und die Web Ontology Language (OWL) zum Einsatz kommen. Als Abfragesprache für die in der Ontologie gespeicherten Daten verwenden wir SPARQL.
Im RDF werden alle Zusammenhänge als Tripel (subjekt, prädikat, objekt) beschrieben. Dabei sind die Subjekte eindeutig bezeichnete Ressourcen (daher mittels des WWW zugängliche Dokumente) bzw. nicht derartige zugängliche Objekte (z.B. Professoren). Prädikate oder Eigenschaften sind binäre Relationen, die zwischen einem Subjekt und einem Objekt bestehen und das Subjekt beschreiben (z.B. „lehrt an“ oder „geboren am“). Objekte können Subjekte (Ressourcen) oder Literale (Werte) sein (z.B. die Ressource „Medizinische Fakultät“ oder das Literal „14. April 1879“). Außerdem kann das Objekt auch eine Variable sein, die für eine Ressource mit beliebigen Eigenschaften steht. Durch sogenannte Reifikation können RDF-Tripel wieder Subjekte anderer RDF-Tripel sein (d.h. es können Aussagen über Aussagen getroffen werden). Weiterhin bietet RDF auch drei verschiedene Typen von Datensammlungen als Datentypen an (ungeordnete und sortierte Listen sowie Alternativlisten). Es existiert eine XML-Syntax für RDF (obschon RDF nicht von der Darstellung als XML abhängig ist), die auch in diesem Projekt Anwendung finden soll.

Als Erweiterung zu RDF existiert die Schema-Sprache *RDF Schema* (RDFS). Mit ihr wird das Vokabular für die Interpretierung der in RDF formulierten Aussagen definiert. RDFS ermöglicht festzulegen, welche Eigenschaften auf welche Art von Objekten anwendbar ist, und welche Werte sie annehmen können (für Eigenschaften wird eine Subklasse Constraint definiert, die hauptsächlich die Instanzen range und domain ermöglichen, d.h. „Definitions- und Wertebereich“). Außerdem wirden die Beziehungen zwischen den Objekten (instanceOf und subClassOf) definiert.

Eine weitere Ergänzung zu RDF / RDFS stellt die *Web Ontology Language* (OWL) dar. Sie ermöglicht eine semantische Erweiterung, indem noch komplexere Bedingungen für den Aufbau von Klassen definiert werden können (z.B. Einschränkun der Zahl und Typen von Eigenschaften einer Klasse). In unserem Projekt werden keine von den fortgeschrittenen Konzepten der OWL angewandt, weshalb wir hier auf eine genauere Beschreibung verzichten möchten.

SPARQL ist ein Protokoll und eine Abfragesprache für das Semantic Web. SPARQL ist ein rekursives Akronym und steht für *SPARQL Protocol and RDF Query Language*. Im März 2007 wurde von der W3C der Last Call Working Draft zur Spezifikation SPARQL vorgestellt.
SPARQL ist für Abfragen im Projekt Professorendatenbank besonders geeignet. Die Syntax ist der Simple Query Language SQL ähnlich, welche seit Jahren bewährt für Datenbankabfragen herkömmlicher Datenbanken eingesetzt wird. Kommen bei SQL meist „SELECT ... FROM ... WHERE ...“-Statements zum Einsatz, so lautet das Pendant in SPARQL „SELECT ... WHERE ...“.

Ein kurzes Beispiel zur Veranschaulichung:

```
PREFIX abc:<http://domain.de/beispielOntologie#>
SELECT ?professor
WHERE {
        ?x abc:professorname "Müller".
        ?x abc:geborenIn "1890"
}
```

Das Beispiel bewirkt auf eine Ontologie, die wie in unserem Projekt einen Katalog von Professoren modelliert, das alle Professoren mit Name „Müller“ und Geburtsjahr „1890“ gesucht werden. Das Schlüsselwort Prefix dient nur der Abkürzung des Pfades der Ontologie, der sonst jedem Prädikat vorangestellt werden müsste. An diesem Beispiel ist sehr schön zu sehen, wie man Suchanfragen in SPARQL formulieren kann. Für professionelle Nutzer wäre es sogar denkbar ein Suchfeld für direkte Formulierung dieser Art von Anfragen anzubieten, aber SPARQL ist auch für Suchanfragen über herkömmliche Eingabemasken einzusetzen.
Es exisitieren Implementierungen in den üblichen Programmiersprachen, für PHP ist beispielsweise RAP (RDF API for PHP; *siehe Rahmenapplikationen* mit SPARQL Client Library und SPARQL Query Engine eine Möglichkeit, SPARQL auf RDF Datastores anzuwenden.

PHP Hypertext Preprocessor ist eine Open-Source Skriptsprache, die hauptsächlich zur Erstellung dynamischer Webseiten oder -anwendungen verwand wird. Für dieses Anwendungsgebiet und damit für unser Projekt ist sie dank großer Internet-Protokollnähe und der hohen Anzahl an verwendbaren Funktions- und Programmbibliotheken besonders geeignet. Das Projekt soll in objektorientiertem PHP erstellt werden. Die Objektorientierung ist eine der jüngeren Erweiterungen von PHP.

In *PHPUnit2* (dem neuen Entwicklungsarm von PHPUnit seit PHP5) wurde eine Testumgebung für PHP-Unittests wie sie für Java mit JUnit 3.8 möglich sind nachempfunden. Damit können schon kleine Code-Fragmente während der Entwicklung auf korrekte Funktion überprüft werden, um eine optimale Validität und Qualität des erzeugten Codes zu jedem Zeitpunkt der Entwicklung zu gewährleisten. PHPUnit unterstützt weiterhin bei der Erstellung der Testdokumentation und kann für Performancetests verwendet werden.

AJAX steht für „Asynchronous JavaScript and XML“ und beschreibt ein Konzept zur asynchronen Übertragung von HTML-Fragmenten über das zustandslose HTTP-Protokoll. Damit ist es möglich bei einer Nutzerinteraktion (beispielsweise einem Klick auf einen Link) nicht eine komplette Seite neu laden zu müssen, sondern nur nach Bedarf entsprechende Teile davon. Es lassen sich somit desktopartige Anwendung auch über das Web realisieren. Unter Verwendung dieser Technik gestaltet sich die Benutzerinteraktion erheblich angenehmer, weshalb wir versuchen werden, sie (bzw. ein entsprechendes Framework) einzusetzen.

Zur Umsetzung der Buchdruckfunktion soll TEX zum Einsatz kommen. TeX ist ein von Donald E. Knuth bereits in den 80er Jahren entwickeltes Textsatzsystem. Im Gegensatz zu den meisten anderen Systemen wie beispielsweise Microsoft Word oder auch Open Office, die das Konzept des „What You See Is What You Get“ verfolgen, werden TeX-Dokumente zunächst in unformatiertem Text geschrieben. Formatiert man ein Dokument in den WYSIWYG-Textsatzsystemen beispielsweise über Menüs und Tastenkürzel, so erfolgt dies bei TeX über eine eigene Syntax. Was zunächst sehr kompliziert und wenig effektiv wirkt, stellt sich nach einer gewissen Einarbeitung als ein erheblicher Vorteil heraus. In diesem Projekt soll die leichte automatische Generierbarkeit von TeX-Code bei herausragenden Ergebnissen genutzt werden, um die gewünschte Buchdruckfunktion so ansprechend wie möglich umzusetzen.

3. Rahmenapplikationen

Die von unserer Gruppe entwickelte Applikation ist kein Plugin oder Modul zu einer bestehenden Rahmenapplikation. Daher werden hier die verwendeteten API, IDE und Framework vorgestellt.

Die *RDF API für PHP* (RAP) ist ein semantisches Webtoolkit für PHP-Entwickler. RAP kam erstmals 2002 als Open-Source-Projekt an der Freien Universität Berlin heraus und wurde seitdem mit internen und externen Codevorlagen erweitert.
Die aktuelle Version beinhaltet unter anderem:

- eine aussagenzentrierte API, um RDF-Graphen mit Hilfe einer Menge von Aussagen zu manipulieren
- eine ressourcenzentrierte API, um RDF-Graphen mit Hilfe einer Menge von Ressourcen zu manipulieren
- einen integrierten RDF/XML, N3 und N-TRIPLE Parser
- einen integrierten RDF/XML, N3 und N-TRIPLE Serializer
- in-memory oder Datenbankmodel-Speicher
- Unterstützung für allgemeine Vokabeln

Das *Zend Framework* ist ein komplett objektorientertes Framework auf PHP-Basis zur Entwicklung von Webapplikationen. Es stellt Komponenten zur Entwicklung anhand des Model-View-Controller-Entwurfsmusters bereit, sowie zur Realisierung von Sessions, Lokalisierung, Suche, Caching, etc. Das Framework folgt dem Ansatz "Don't Repeat Yourselfünd stellt somit immer wiederkehrende Aspekte von Webanwendungen zur Verfügung, die sehr häufig benötigt werden.

Unsere Projektgruppe wird zum Erstellen des PHP-Codes die Open-Source Entwicklungsumgebung *Eclipse* verwendet. Mit dem PHPeclipse-Plugin wird Eclipse für die Verarbeitung von PHP vorbereitet. Eclipse ermöglicht weiterhin einen integrierten CVS-Zugriff auf die Team-Ressourcen.
Mit dem Tool ArgoUML (auch als Eclipse-Plugin argoeclipse) kann sowohl die Modellierung des Projektes in UML vorgenommen werden, als auch die PHP-Funktionsrümpfe generiert werden.