Dateiformate für das Elektronische Publizieren

Von Daniel Ohst, Projekt „Digitale Dissertationen“, Humboldt-Universität zu Berlin / Rechenzentrum
E-Mail: ohst@informatik.hu-berlin.de

Einleitung

Das Thema „Dateiformate für das elektronische Publizieren“ ist ein Dauerbrenner. Obwohl sich schon seit geraumer Zeit im Rahmen von zahlreichen Forschungsprojekten diesem Gebiet gewidmet wird, gibt es bis heute kein Patentrezept oder allgemeingültige Regeln für die Verwendung bestimmter Dateiformate für die Publikation digitaler Dokumente. Die Gründe dafür liegen sowohl in der Heterogenität der Hardware- und Softwarelandschaft, den unterschiedlichen Anforderungen der Autoren an Textsysteme aber auch an der Betrachtung des elektronische Publizierens auf unterschiedlich komplexen Ebenen.
Neben einer Übersicht zu den aktuell am meisten verwendeten Dateiformaten, soll der Ansatz des Projekts „Digitale Dissertationen“ – die Verwendung von SGML als Archivierungs- und Rechercheformat – vorgestellt werden. Dabei sollen sowohl die Vorteile dieser Herangehensweise präsentiert aber auch die Nachteile nicht verschwiegen werden.

Formate

Mit welchen Formaten hat man es zu tun, wenn man heute im wissenschaftlichen Umfeld elektronisch publizieren möchte?
Für die Erstellung von Dokumenten werden hauptsächlich die Formate MS-Word und TeX (LaTeX) benutzt. Weitere, wie z.B. Starwriter, Framemaker oder Wordperfect, spielen derzeit eine untergeordnete Rolle.
Im WWW werden Dokumente in den Formaten HTML bzw. PDF publiziert. Da HTML trotz der intensiven Entwicklung im Bereich Style-Sheets noch nicht über ausgefeilte Layoutmöglichkeiten verfügt, werden vielfach das Portable Document Format eingesetzt, das als layoutorientiertes Format, ähnlich wie Postscript, hier HTML überlegen ist, aber z.B. einen eigenen Viewer zur Darstellung im Browser benötigt.
Die Frage der Verwendung von Dateiformaten für die Archivierung von Dokumenten ist weit schwieriger zu beantworten. Wenn überhaupt eine professionelle Archivierung stattfindet, wird vielfach das Format gespeichert, das auch als Präsentationsformat dient. Ob dieses allerdings in Zukunft problemlos lesbar sein wird, läßt sich nicht zweifelsfrei beantworten.

Der "DiDi-Ansatz“

Das Projekt hat sich dafür entschieden, das Format SGML für die Recherche und die Archivierung der digitalen Dissertationen zu verwenden. SGML ist ein ISO-Standard, der eine Sprache definiert, mit der Dokumentstrukturen abgebildet werden können, d.h. für jede Klasse gleichartiger Dokumente, z.B. Dissertationen, wird eine Document Type Definition (DTD) erstellt, die beschreibt, aus welchen Teilelementen der Text besteht. In gewisser Weise lassen sich diese DTDs mit Formatvorlagen in Word vergleichen, allerdings mit der wesentlichsten Einschränkung, daß in SGML keinerlei Layout definiert wird. Dies wird vielmehr in einer separaten Styledefinition vorgenommen.
Speziell für den Anwendungsfall Dissertationen wurde im Rahmen des Projekts die Dokumenttypdefinition „DiML“ entwickelt, die auf einer DTD der VirginiaTech-Initiative beruht. Es wird jedoch kein Promovend gezwungen, nun seine Dissertation in SGML schreiben zu müssen, dies wird vielmehr die Ausnahme bleiben. Zu ca. 80% wird an der HU mit MS Word gearbeitet. Aus diesem Grunde wurde eine Formatvorlage entwickelt, in der sich die DTD „DiML“ gut abbilden läßt. Diese wird den Promovenden zur Verfügung gestellt. Darüberhinaus werden Onlinehilfen zur Benutzung und regelmäßig Kurse angeboten, wo die Promovenden den praktischen Umgang mit

der Formatvorlage erlernen können. Die Erfahrungen und Feedbacks aus diesen Kursen sind äußerst positiv.
Ein mit dieser Formatvorlage erstelltes Dokument wird dann unter Verwendung von Konvertierungswerkzeugen (z.B. MS SGML Author, Wordperfect, Omnimark) nach SGML umgewandelt. Da eine direkte Publikation von SGML im WWW derzeit aufgrund der fehlenden Browserunterstützung nicht möglich ist, wird eine HTML- und eine PDF-Version erzeugt und auf dem Dokumentenserver abgelegt. Die Recherche erfolgt jedoch direkt über dem SGML-Dokument, so daß alle Vorteile, wie z.B. Suche in Strukturen, genutzt werden können. Für die Archivierung werden derzeit SGML und die dazugehörigen Styles sowie das PDF-Dokument verwendet.

SGML – Vorteile und Nachteile

Der Einsatz von SGML heißt, auf einen ISO-Standard zu setzen, der bereits seit geraumer Zeit im Einsatz ist. Die Entwicklung und Nutzung einer DTD bedeutet, eine einheitliche, standardisierte Beschreibung einer Menge von Dokumenten vorzunehmen, was die spätere Weiterverarbeitung wesentlich vereinfacht. Mit SGML wird die Struktur der Originaldokumente beibehalten und nicht wie bei Layoutformaten wie PDF vernichtet. Dadurch lassen sich SGML-Dokumente später flexibel in andere Formate umwandeln, z.B. für die Publikation in HTML.
Bei der Archivierung spielt dieses Kriterium eine entscheidende Rolle.
Des weiteren ermöglicht die Strukturierung von SGML neue Arten von Suchanfragen bei der Recherche. So lassen sich z.B. ganz gezielt Orte oder Gutachter über einer Menge von Dokumenten recherchieren.

Um jedoch überhaupt SGML für die Publikation in Betracht ziehen zu können, müssen natürlich auch strukturierte Texte vorliegen. Falls z.B. eine Überschrift in einem Word-Dokument statt der Formatvorlage Überschrift einfach eine größere Schriftart zugewiesen bekommen hat, ist dies natürlich keine Strukturierung des Textes und somit zwar in SGML darstellbar, allerdings ohne die obengenannten Vorteile. Es ist also von großer Bedeutung, den Autoren die Vorteile des strukturierten Schreibens von Dokumenten klar vor Augen zu führen.
Weitere Probleme sind eher technischer Art: Die komplexen und vergleichsweise teuren Tools sind derzeit noch nicht in der Lage, den Konvertierungsprozeß weitgehend automatisch ablaufen zu lassen. So gibt es z.B. Probleme bei komplizierten Tabellen, und auch die Seitenidentität zwischen Papier-, PDF- und SGML-Version läßt sich derzeit nicht ohne manuelle Eingriffe herstellen.
Für eine direkte Publikation des gewonnenen SGML-Textes im WWW fehlt es derzeit noch an der notwendigen Browserunterstützung. Es müssen also noch weitere, allerdings vollautomatische, Konvertierungen z.B. nach HTML ablaufen.

Fazit

SGML ist das vom Projekt favorisierte Format für Recherche und Archivierung digitaler Dissertationen. Durch die Trennung von Struktur und Layout eines Dokuments ergeben sich neue Möglichkeiten der Recherche, die durch eine Volltextsuche nicht repräsentierbar sind. Für die Archivierung erhält SGML sämtliche Strukturinformationen des Originaltextes und bietet somit sehr gute Voraussetzungen für eine spätere Konvertierung in zukünftige Formate.
Der Aufwand für die Publikation in SGML ist sicherlich derzeit wesentlich höher anzusetzen als für die Erzeugung eines PDF-Dokuments. Ziel muß es sein, den Zeitaufwand und die Komplexität des gesamten Prozesses soweit zu reduzieren, daß er im Tagesgeschäft der Bibliothek integriert werden kann, was im Projekt zielstrebig verfolgt wird.
Die Nachteile des SGML-Ansatzes wurden im vorangegangenen Abschnitt angerissen. Mit der weiteren Entwicklung von XML wird sich jedoch auch die Browserunterstützung verbessern, so daß eine direkte Publikation im WWW möglich wird. Auch die Tools für die Konvertierung werden ständig erweitert. Somit ist zu erwarten, daß hier in naher Zukunft die meisten der technischen

Probleme beseitigt sein werden. Wesentlich ist jedoch weiterhin der Schulungsbedarf im Bereich strukturiertes Schreiben von Texten. Dieses Problem läßt sich nicht durch technische Hilfsmittel lösen, sondern muß in den Köpfen der Schreibenden permanent präsent sein. Die Publikation in SGML macht keinen Sinn, wenn der zugrundeliegende Text keine Struktur besitzt.