Digitale Dissertation

# Die „Humboldt-Dissertationen“ im Internet

Erste Ergebnisse und Erfahrungen aus dem Projekt DiDi

## Elektronische Dissertationen gemäß Promotionsordung

Seit kurzem sind sie offiziell möglich, die digitalen Dissertationen an der Humboldt-Universität. Mit dem Beschluß des Akademischen Senats vom 24. Februar 1998 wurde der Grundstein für das elektronische Publizieren von Doktorarbeiten an dieser Universität gelegt. Dadurch wurde die Internetpublikation auf dem offiziellen Dokumentenserver der HU als eine Möglichkeit etabliert, der Veröffentlichungspflicht für Doktorarbeiten nachzukommen. Ein entsprechender Zusatz zu sämtlichen Promotionsordnungen der Humboldt-Universität fand Zustimmung beim Akademischen Senat und wurde von diesem einstimmig beschlossen. Diese Tatsache bildet die Grundlage für die zukünftige Arbeit des Projektes „Digitale Dissertationen“.

## Das Gesamtkonzept

Zu Beginn des Projektes wurden die verschiedenen Anforderungen, die ein Konzept zur elektronischen Publikation von Dissertationen beinhalten sollte, definiert [1]. Als Hauptpunkte wurden rechtliche, bibliothekarische [2] und rechentechnische Anforderungen aufgestellt, auf deren Grundlage sich im wesentlichen drei große Aufgabengebiete für das Projekt ergeben: 1. die Festlegung von Dateiformaten für die Abgabe, Präsentation, Archivierung und Recherche sowie die Entwicklung von Konvertierungswerkzeugen, 2. die Installation eines Dokumentenservers, inklusive Konzept zur Sicherung der Authentizität der eingespielten Dokumente und 3. die technische Beratung und Betreuung der Promovenden bei der Erstellung der Arbeiten.

## Dateiformate

Für die Entwicklung eines Konzepts für einen Dokumentenserver spielt die Frage nach der Wahl der Dateiformate [3] eine zentrale Rolle. Zum einen sollte die Darstellung eines elektronischen Dokuments im Internet der abgegebenen Druckversion entsprechen, um die Zitierfähigkeit des digitalen Exemplars zu gewährleisten. Ein großes Problem bildet dabei z. B. die Übernahme der Seitenzahlen in das elektronische Dokument. Zum anderen wird vom Dateiformat gefordert, daß es eine langfristige Lösung zur Speicherung der elektronischen Dokumente und für deren strukturierte Recherche darstellt. Wünschenswert wäre weiterhin, daß die Erstellung des Dokuments vom Autor in seinem gewohnten Textverarbeitungssystem vorgenommen werden kann. Diese Forderungen führen zu der Überlegung, zwischen Abgabe-, Archivierungs-, Recherche- und Präsentationsformaten zu unterscheiden.

Zulässige Abgabeformate für digitale Dissertationen sind die Formate RTF (Rich Text Format) und LaTeX. RTF kann aus allen gängigen Textverarbeitungsprogrammen der Windows-Welt, wie z. B. WinWord oder WordPerfect erzeugt werden, ohne daß komplizierte Konvertierungsmechanismen benutzt werden müssen. LaTeX wird als ein äußerst leistungsfähiges Textsatzsystem vor allem in den Natur- und Ingenieurwissenschaften (z. B. Mathematik, Physik, Chemie, Informatik) eingesetzt, da seine Fähigkeiten zur Darstellung von mathematischen Formeln die anderer Systeme weit übertreffen. Es muß daher seitens des Projektes eine Unterstützung dieser Klientel gewährleistet werden.

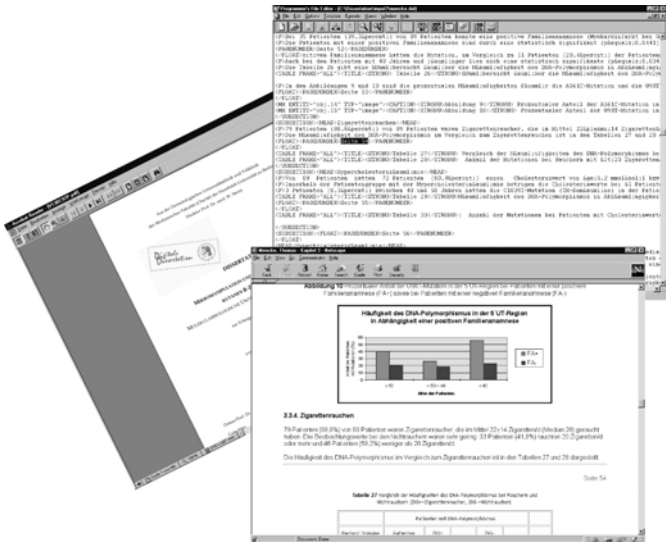

Abb. 1: Elektronische Dissertation: PDF-SGML (Sourcecode) - HTML (Browserdarstellung)

Als Archivierungs- und Rechercheformat findet SGML Anwendung. Die **S**tandard **G**eneralized **M**arkup **L**anguage (SGML) wurde 1969 aus der Generalized Markup Language (GML), einem IBM-Produkt, entwickelt. Sie ist seit 1986 ein ISO-Standard (ISO 8879). Lange Zeit lag ihr Einsatzgebiet vorrangig im Bereich der technischen Dokumentation, und erst Anfang der neunziger Jahre begannen Verlage, diese Publikationsmethode für sich zu entdecken. SGML kann, streng genommen, nicht als richtiges Dateiformat bezeichnet werden. Es ist eher ein Konzept, welches darauf ausgelegt ist, den Inhalt, d. h. den eigentlichen Text eines Dokumentes, von seiner logischen Struktur und dem Layout (Markup) zu trennen. Um ein Dokument zu erstellen, muß zunächst seine Struktur definiert

werden. Dies wird in einer sogenannten Document Type Definition (DTD) getan, die jeweils für eine Klasse gleichartiger Dokumente steht, in unserem Beispiel Dissertationen. In einer DTD werden sowohl die logischen Elemente, wie z. B. Überschriften, Absätze, Fußnoten, Zitate definiert, als auch der Kontext und die Anzahl ihres Auftretens festgelegt. Im eigentlichen SGML-Dokument wird der Text in die definierten Elemente (Tags) geschrieben. Durch seine definierte Struktur ist SGML besonders gut für die Recherche in Dokumenten gleichen Typs geeignet. Es zeichnet sich auch dadurch aus, daß es vor allem aufgrund der Unabhängigkeit von einzelnen Anbietern das Format ist, das die größte Gewähr für die Lesbarkeit auch in künftigen Jahrzehnten bietet und somit für eine Archivierung geeignet ist.

Die Präsentation der Dissertationen erfolgt in den Formaten PDF (**P**ortable **D**ocument **F**ormat) und HTML (**H**yper**T**ext **M**arkup **L**anguage). Mit Hilfe des PDF-Formates, welches mit dem Adobe Acrobat-Paket aus der RTF-Version bzw. über den Umweg Postscript aus der LaTeX-Version erzeugt wird, erreicht man eine dem Druckexemplar fast identische Darstellung der Arbeit im Internet. HTML ist das heute im Internet gebräuchliche Textformat und kann mit jedem Browser gelesen werden. PDF dagegen bietet die Gewähr für ein einheitliches Layout und kann eine Druckausgabe am ehesten simulieren. Der Doktorand hat also die Sicherheit, daß sein Dokument formal ansprechend und auf allen Rechnern in gleicher Weise erscheint.

HTML, welches vom World Wide Web Consortium standardisiert wurde, liegt aktuell in der Version 4.0 vor und kann als eine Dokumenttypdefinition im Sinne von SGML angesehen werden. Da es sich im WWW durchgesetzt hat, gibt es für beliebige Plattformen umfangreiche und frei verfügbare Tools zur Erstellung, Verwaltung, Konvertierung und Betrachtung von HTML-Dokumenten. Der Nachteil liegt darin, daß die Darstellung mathematischer Symbole und Formeln nicht unterstützt wird. Daher wird längerfristig auf Formate wie XML (Extensible Markup Language) und seine Instanz MathML (Mathematical Markup Language) orientiert.

Das PDF-Format stellt eine Weiterentwicklung des De-facto-Standards Postscript dar und ist wie dieses eine Seitenbeschreibungssprache. Im Unterschied zu Postscript ist PDF für das Publizieren im WWW konzipiert. Um PDF-Dokumente im Internet betrachten zu können, ist es für den Benutzer erforderlich, daß er vorher auf seinem WWW-Browser, z. B. Netscape Communicator oder Microsoft Internet Explorer, das Acrobat Reader Plugin installiert hat, welches die Firma Adobe kostenfrei zur Verfügung stellt.

## Dokumentenserver

Ziel ist es, einen Server zu installieren, der sich von herkömmlichen WWW-Servern dadurch unterscheidet, daß hier nur von der Humboldt-Universität zugelassene Dokumente eingespielt werden, deren Authentizität und Integrität gewährleistet werden müssen. Erster Schritt zur Installation eines zertifizierten[1] Dokumentenservers (Adresse: http://dochost.rz.hu-berlin.de/docserv/) war die Auswahl eines Datenbanksystems zur Speicherung der elektronischen Volltexte. Hier nutzen wir für eine Übergangsperiode den Hyperwave Information Server. Vorteil dieses Produktes ist die Tatsache, daß URLs (Uniform Resource Locator) von Dokumenten innerhalb des Systems konstant gehalten werden, selbst wenn es zur Verschiebung von einzelnen Elementen kommt. Hyperwave unterstützt eine Volltextrecherche in den unterschiedlichsten Sammlungen auf seinem Server. Auf diese Weise werden die PDF- und die HTML-Versionen der Arbeiten organisiert und recherchiert. Da hier das Struktur-Format SGML nicht explizit unterstützt wird, muß auf eine Doppellösung zurückgegriffen werden, möchte man den Vorteil der strukturierten Suche in SGML-Dokumenten effizient ausnutzen. Hier bietet sich das Harvest-System an, da es in der Lage ist, bei entsprechender Konfiguration des Gatherers und des Brokers unterschiedliche DTDs von SGML zu unterstützen.

Die aktuelle Lösung (siehe Abbildung 2) sieht demnach so aus, daß für die Volltext- und die strukturierte Recherche in den SGML-Dokumenten Harvest benutzt wird, bei der Präsentation der Rechercheergebnisse jedoch auf die HTML-Varianten zurückgegriffen wird. Eine Recherche über die PDF- oder die HTML-Dokumente wird durch die vom Hyperwave Server angebotene Suche (eine Verity Search Engine) realisiert. In Vorbereitung ist momentan die Ausschreibung eines Datenbanksystems zur strukturierten Suche in SGML-Dokumenten.

Ein weiterer wesentlicher Punkt in der Entwicklungsarbeit wird die Anbindung des Dokumentenservers an das Bibliothekssystem über eine Z39.50-Schnittstelle sein. Dies wird es dem Benutzer ermöglichen, über den OPAC (Online Public Access Catalogue) der Universitätsbibliothek auf die elektronischen Volltexte direkt zuzugreifen. Hierzu muß eine teilweise automatisierte Erfassung von Metadaten etwa nach dem Dublin Core-Standard [4] realisiert werden.

Um eine dauerhafte Archivierung gewährleisten zu können, ist eine Integritätssicherung des Dokumentenservers und der abgelegten Dokumente unabdingbar. Bei den elektronischen Dissertationen müssen Autor, Inhalt und Veröffentlichungszeitpunkt der Dokumente beweiskräftig vor Fälschung und Zweifeln an der Authentizität geschützt werden. Dazu werden in Zusammenarbeit mit dem HU-Projekt „Firewall – ein Kernstück zur Sicherung des Verwaltungsnetzes“ Konzep-

1 Dem Dokumentenserver muß von der Zertifizierungsstelle der Humboldt-Universität, der HU-CA (http://ca.hu-berlin.de/) ein Signaturschlüsselzertifikat ausgestellt werden.

te zu digitalen Signaturen und Zeitstempelverfahren erprobt. Ein Dokument kann nur von berechtigten Personen auf den Hyperwave-Server eingespielt werden, da ein Paßwortschutz für definierte Gruppen und Personen besteht. Eine digitale Signierung ist für die PDF-Version der jeweiligen Arbeit als Abbild des Druckexemplars vorgesehen. Dazu ist die Ausstellung eines Signaturschlüsselzertifikats für den Dokumentenserver durch eine Zertifizierungsinstanz notwendig. Dieses Doktoranden die Recherchemöglichkeiten des Internets für ihre wissenschaftliche Arbeit nutzen und inwieweit sie bereit sind, selbst im Internet zu veröffentlichen. Wichtig war es auch, den Kenntnisstand in Bezug auf das strukturierte Schreiben und die Nutzung von Formatvorlagen zu erfassen. Im Ergebnis der Befragung wurde festgestellt, daß Promovenden vor allem in WinWord ihre Arbeiten verfassen und sich in der Verwendung von Formatvorlagen bzw. Struktur noch

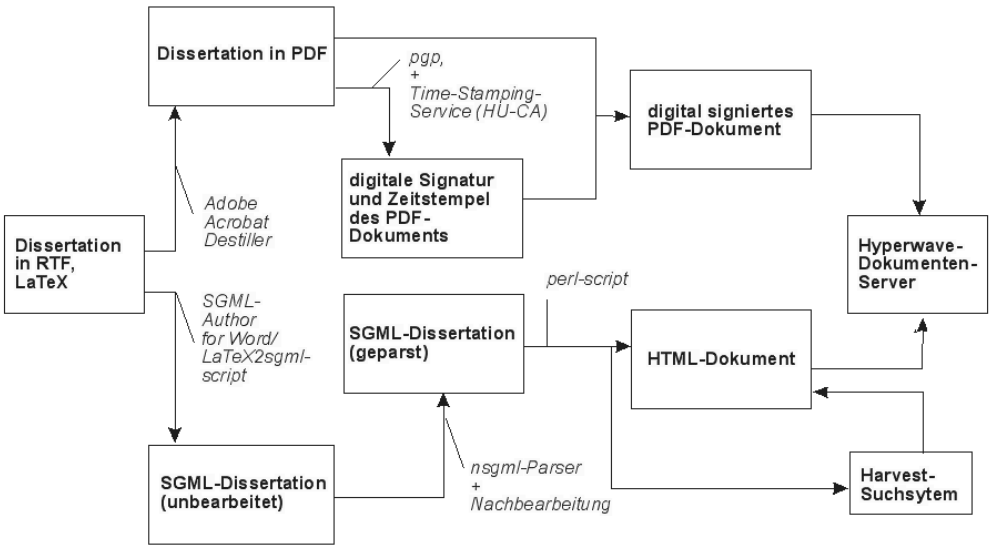


Abb. 2: Workflow einer Technologiestudie „Digitale Dissertationen“

Schlüsselzertifikat ordnet dem Dokumentenserver eindeutig seinen öffentlichen Schlüssel zu (vgl. auch [5]). Für jede eingespielte PDF-Dissertation wird eine Prüfsumme gebildet, die vom Dokumentenserver digital signiert wird. Das kann z. B. durch das PGP-Verfahren realisiert werden (siehe Artikel „*Vertrauen gegen Vertrauen*“ von Alexander Geschonneck in diesem Heft). Zusätzlich wird ein digitaler Zeitstempel erzeugt. Ein Zeitstempel im Sinne des SigG [6] ist eine mit einer digitalen Signatur versehene digitale Bescheinigung einer Zertifizierungsstelle, daß ihr bestimmte digitale Daten zu einem bestimmten Zeitpunkt vorgelegen haben [7].

## Autorenbetreuung

Dem dritten Schwerpunkt des Projektes haben wir bisher die größte Aufmerksamkeit gewidmet. Ausgangspunkt war eine Befragung potentieller Promovenden der Humboldt-Universität durch einen Fragebogen. Dieser wurde sowohl in Papierform über die Zweigbibliotheken der Universitätsbibliothek verteilt, als auch über die WWW-Seiten des Projektes zum Ausfüllen zur Verfügung gestellt, und die Doktoranden wurden über eine Rundmail von dieser Befragung in Kenntnis gesetzt. Ziel war es, in Erfahrung zu bringen, inwieweit ein großer Schulungsbedarf ergibt. Daher stand zunächst die Umsetzung einer ersten für die Humboldt-Dissertationen anzuwendenden DTD[2] in eine unter WinWord zu verwendende Formatvorlage im Vordergrund. Sinn dieser Umsetzung ist, daß die Doktoranden weiter in ihrem gewohnten Textverarbeitungssystem schreiben, die Umsetzung in das strukturierte SGML-Format jedoch weitgehend automatisiert geschehen kann. Zusätzlich wurde ein Windows-Hilfesystem für die Formatvorlage entwickelt, welches aktuell in ein Hypertexthilfesystem überführt wird, um über die WWW-Seiten abgerufen zu werden. Die Information über die Möglichkeit, seine eigene Dissertation über den Dokumentenserver der Humboldt-Universität im WWW zur Verfügung zu stellen, muß an der Universität noch weiter verbreitet werden. Aus diesem Grund führen die Mitarbeiter des Projektes regelmäßig Informationsveranstaltungen in den Fakultäten durch, wobei im Moment der Fokus auf die Gewinnung von Promovenden der Medizinischen Fakultät Charité gelegt wurde. Aus den Diskussionen mit interessierten Doktoranden haben wir einen speziellen Kurs zur Einführung in das Schreiben von Doktorarbeiten in WinWord mit unserer Formatvorlage konzipiert und bereits erprobt. Nächste Schritte sollten sich mit der Nutzung der SGML-Formatvorlage in WordPerfect und vor allem mit der Entwicklung von Vorlagen für die LaTeX-Nut-

2 Die DTD heißt DiML (Dissertations Markup Language)

zer beschäftigen, da Mitglieder der naturwissenschaftlichen Fakultäten traditionell aufgeschlossener solchen Dingen wie der Internetpublikation von Doktorarbeiten gegenüberstehen. Im Moment können wir hier noch keine qualitativ hochwertige Unterstützung bei der Erstellung und Umsetzung der Arbeiten in unsere DiML anbieten. Auch scheint es notwendig zu sein, zwischen den einzelnen Fachgruppen stärker zu differenzieren und so fachspezifische DTDs zu entwickeln. Ein großes Problem bildet immer noch die Akzeptanz des gesamten Anliegens vor allem in den geisteswissenschaftlichen Fächern. Einen Höhepunkt in der Information der Universität zum Thema Elektronisches Publizieren von Dissertationen wird das öffentliche Kolloquium des Rechenzentrums am 10. Juni 1998 bilden.

## Initiativen an anderen Universitäten

DiDi steht nicht allein da. Im bundesweiten Umfeld gibt es an fast jeder Universität Bemühungen, die den Aufbau eines Volltextarchivs vorsehen. Eine Koordinationsfunktion soll hierbei das Projekt „Dissertationen-Online“[8] übernehmen. Das Projekt der IuK-Kommission von sechs wissenschaftlichen Fachgesellschaften (Chemie, Physik, Mathematik, Erziehungswissenschaften, Soziologie und Informatik) ist an verschiedenen Universitäten der Bundesrepublik angesiedelt. Gesamtanliegen dieses Projektes ist es, Kontakte zu den Fakultäten und Bibliotheken aufzunehmen, fachtypische Dissertationen zu analysieren und daraus Strukturen für elektronische Dissertationen zu entwickeln, Promovenden bei der Erstellung derartiger Arbeiten beratend zur Seite zu stehen und zusammen mit den Bibliotheken und Rechenzentren die technische Umsetzung zu leisten. Das Teilprojekt der Autorenbetreuung, geleitet von Prof. Diepold (Abteilung Pädagogik und Informatik am Institut für Wirtschafts- und Erwachsenenpädagogik, Philosophische Fakultät IV), wird an der Humboldt-Universität durchgeführt.

Vorbild für das gesamte Vorhaben bilden verschiedene Initiativen amerikanischer Universitäten. Besonders hervorzuheben ist in diesem Zusammenhang das „Electronic Thesis and Dissertations Project“ (ETD) der University of Virginia [9].

## Fazit

Von den Forderungen und Konzepten, die im April 1997 in dem Artikel von P. Schirmbacher und N. Martin formuliert wurden, ist schon einiges umgesetzt worden. Unsere Erfahrungen in den ersten Monaten haben gezeigt, daß die Information und Betreuung der Autoren einen ganz wesentlichen Teil der weiteren Projektarbeit bilden muß.

Fragen wie die der Akzeptanz elektronischer Publikationen im Vergleich zu Verlagspublikationen oder die des eigenen Zitiernachweises (citation index) führen in den Informationsveranstaltungen häufig zu intensiven Diskussionen über Urheberrechte und Publikationsstrategien.

Um die hohe Qualität der Recherche und der Langzeitarchivierung der digitalen Dissertationen gewährleisten zu können, muß ein Teil der Verantwortung an die Promovenden als die Autoren der digitalen Publikationen weitergegeben werden. Das betrifft besonders die Einarbeitung von Strukturinformationen in die Dokumente. Dies bildet die Grundlage für eine erfolgreiche Überführung der Dissertationen nach SGML und damit ihrer Recherchierbarkeit. Hier wird der große Schulungsbedarf offensichtlich, den die Promovenden als Autoren digitaler Publikationen haben.

Susanne Dobratz
susanne.dobratz@rz.hu-berlin.de

**Literatur**

[1] Martin, N., und Schirmbacher, P.: Die elektronische Publikation von Dissertationen. RZ-Mitteilungen, 14 (1997).
[2] Siehe dazu auch Martin ‚N.: Die elektronische Publikation von Dissertationen – Bibliothekarische Anforderungen. WWW-Seiten des Projektes: http://dochost.rz.hu-berlin.de/epdiss/projekt.html
[3] Ohst, D.: Dateiformate für das elektronische Publizieren. Studienarbeit, Institut für Informatik, Humboldt-Universität zu Berlin, 1998, http://www2.rz.hu-berlin.de/ ~h0444saa/didi/formate.html
[4] Siehe http://www.oclc.org:5046/research/dublin_core/
[5] Ohst, D.: Verschlüsselung im WWW. RZ-Mitteilungen, 15 (1997), 22 - 26.
[6] Gesetz zur digitalen Signatur – Signaturgesetz (SigG) vom 1. August 1997, Artikel 3 des Informations- und Kommunikationsdienste-Gesetzes (IuKDG), http://www.iid.de/rahmen/inkdgbt.html
[7] vgl. Artikel von Meinholdt, Luckhard in der c’t, 8 (1998).
[8] http://www.educat.hu-berlin.de/diss_online/
[9] http://etd.vt.edu/
[10] Fox, D.: Beweismittel – Unterschriften auf der Datenautobahn, iX, 12 (1997), 98-100.
[11] Meinhold, M.; Luckhard, N.: Echtheits-Zertifikat – Digitale Signaturen mit beweiskräftigem Zeitstempel, c’t, 8 (1998), 112 - 116.
[12] Ohst, D., und Schirmbacher, P.: Zur Wahl von Dateiformaten für die elektronische Publikation von Dissertationen an der Humboldt-Universität zu Berlin. Rechenzentrum der HU, 1996.
[13] Rieger, W.: SGML für die Praxis. Springer Verlag, 1995.
[14] http://medoc.informatik.tu-muenchen.de/