XML-Portal: Archivierung von Multimediaobjekten mit XML-DTDs

Seit Oktober 2001 wird im Rahmen des DFG-geförderten Projektes XML-Portal der Dokumentenserver der Humboldt-Universität, der bisher primär auf die Verwaltung und Archivierung textbasierter Dokumente spezialisiert war, so erweitert, dass auch Multimediadokumente unter Nutzung standardisierter Objektbeschreibungen verwaltet, recherchiert und archiviert werden können.

Ausgangslage

Universitätsbibliotheken, Archive und Museen stehen angesichts der ständig steigenden Zahl elektronischer Publikationen vor dem Problem, für diese adäquate Strategien der Erschließung, des Zuganges und der Archivierung anzubieten. Dabei mangelt es sehr häufig an standardisierten Lösungen und nutzerorientierten Herangehensweisen. Ein Konzept, welches XML (eXtensible Markup Language) als Basis für ein „portal computing“ nutzt, kann als Ausgangspunkt dienen, die Erschließung, Archivierung und Verwaltung – insbesondere auch multimedialer Objekte – unter bibliothekarischen, dokumentarischen und rechentechnischen Gesichtspunkten befriedigend vornehmen zu können.

Im Projekt XML-Portal sollen die Erfahrungen, die an der Humboldt-Universität im Bereich des elektronischen Publizierens für textorientierte Dokumente im Rahmen des HSP III-Projektes „Digitale Dissertationen“[1] , siehe [1], [2], und des DFG-Projektes „Dissertationen Online“[2] gemacht wurden, aufbereitet und auf multimediale Dokumente angewendet werden. Das für textorientierte Dokumente bestehende Konzept umfasst eine nachnutzungsfähige Lösung von der Autorenbetreuung über einen automatisierten Workflow bis hin zum Retrieval.

Im neuen Projekt, gefördert von der DFG[3] , wird das Ziel verfolgt, ein standardisiertes Konzept zu erarbeiten, welches die Aufnahme, Verwaltung und Recherche auch von multimedialen Dokumenten ermöglicht. Dies ist in dem bisherigen Konzept des Dokumentenservers nicht gegeben. Als Beispielanwendung sollen in Zusammenarbeit mit dem Kunstgeschichtlichen Seminar und dem Helmholtz-Zentrum für Kulturtechnik die Sammlungen der Humboldt-Universität genutzt werden. Die Projektpartner Rechenzentrum (Dr. Peter Schirmbacher), Bibliothek (Dr. Norbert Martin) und Institut für Informatik (Prof. Dr. Wolfgang Coy) streben an, eine nachnutzungsfähige und übertragbare Lösung zu schaffen.

Durch die Zusammenarbeit der Humboldt-Universität mit der Staats- und Universitätsbibliothek Göttingen im Rahmen eines anderen Projektes (ProPrint: DFN-Verein)[4], vgl. [6], können die dort erzielten Ergebnisse des Zusammenschlusses zweier heterogener Dokumentenserver ebenfalls in das Projekt einfließen und genutzt werden.

Das Problem: Die standardisierte Erfassung von Medienobjekten

Die Problematik der Speicherung und Verwaltung multimedialer Objekte ist – wie die Praxis zeigt – Gegenstand der Entwicklung technologischer Abläufe nicht allein in Universitätsbibliotheken, sondern tritt heute ebenso in Museen und Kunstarchiven verstärkt zutage. Deshalb orientiert sich das Projekt auch an den Ergebnissen aus Projekten, die sich in den letzten Jahren mit diesem Thema beschäftigt haben, wie z. B.

- Foto Marburg im Kunstgeschichtlichen Institut der Philipps-Universität Marburg,
- MALVINE, ein Projekt der Staatsbibliothek zu Berlin,
- „BAM – Gemeinsames Internetportal für Bibliotheken, Archive und Museen“-Projekt[5] in Baden-Württemberg oder
- Projekte der CIMI-Initiative[6].

Leider fehlt oft die Anwendung von standardisierten Datenformaten zur Objektbeschreibung. In den letzten Jahren hat sich XML (eXtensible Markup Language), eine Untermenge der Standard Generalized Markup Language (SGML), als ein Austausch- und Beschreibungsformat herausgebildet, mit dessen Hilfe eine standardisierte Erfassung multimedialer Objekte und deren Metadaten möglich ist. Es sind so plattform- und herstellerunabhängige Möglichkeiten entstanden, Standardisierungen vorzunehmen. In den ersten Jahren der XML-Entwicklung und der entsprechenden Standardisierungsbestrebungen standen vielfach die textorientierten Anwendungen im Vordergrund. Dies hat sich in der Zwischenzeit geändert, so dass der vermehrte Einsatz von XML auch zur Beschreibung und Präsentation von multimedialen Objekten realistisch ist. In die Nutzung von XML wurde bereits in den RZ-Mitteilungen, vgl. [3] und [4], eingeführt.

1 http://www.edoc.hu-berlin.de/epdiss
2 http://www.dissonline.de
3 Deutsche Forschungsgemeinschaft (http://www.dfg.de). Gefördert wird das Projekt im Rahmen des Programms „Informations-Infrastrukturen für netzbasierte Forschungs kooperation und digitale Publikationen“.
4 siehe Artikel zu ProPrint von Matthias Schulz in diesem Heft oder http://edoc.hu-berlin.de/proprint
5 http://www.bam-bw.de
6 http://www.cimi.org

Ein Projekt, welches sich auf einer allgemeineren Ebene bereits dieser Problematik gewidmet hat und einen ersten Prototypen erarbeiten konnte, ist das Projekt „BAM – Gemeinsames Internetportal für Bibliotheken, Archive und Museen“[7] in Baden-Württemberg. In diesem Projekt wird am Beispiel des Bibliotheksservicezentrums Baden-Württemberg, der Landesarchivdirektion in Stuttgart und des Landesmuseums für Technik und Arbeit in Mannheim ein abgestimmter Dublin Core Metadatensatz, siehe [5], genutzt, um eine gemeinsame Recherchemaske über die Bestände der drei Einrichtungen aufzubauen. Eine Kooperation des Projektes der Humboldt-Universität mit dem Projekt „BAM“ wird angestrebt.

Die Integration verschiedener Bilddatenbanken über XML-Metadatenschnittstellen

An der Humboldt-Universität werden aktuell im Rahmen der Schaffung des interdisziplinären Helmholtz-Zentrums für Kulturtechnik die verschiedenartigen Sammlungen zusammengetragen, digitalisiert und wissenschaftlich aufgearbeitet. Dazu gehören das Lautarchiv, die Porträtsammlung und die medizinische Sammlung Rudolf Virchows. Für jede dieser Sammlungen soll eine auf die Spezifika der jeweiligen Objekte zugeschnittene Datenbank erarbeitet werden. Zum Teil ist dies schon durch das Helmholtz-Zentrum begonnen worden. Ziel ist es, alle Datenbanken später durch die Universitätsbibliothek verwalten und pflegen zu lassen.

Es soll ein definierter Einstiegspunkt zur Recherche in den verschiedenen Informationsangeboten der Bibliothek angeboten werden. Die derzeit erarbeiteten Objektdatenbanken bilden dabei den Grundstock für ein ständig erweiterbares System. Wichtig dabei ist die Schaffung einer Metadatenschnittstelle, welche die Recherche über die verschiedenen Objektdatenbanken gleichzeitig ermöglicht. Diese Schnittstelle wird unter Nutzung von XML-Technologien und entsprechend standardisierter (XML/RDF-)Interpretationen des Dublin Core Metadatenstandards aufgebaut.

Exemplarischer Aufbau eines Bibliotheksportals mit fachspezifischen Sichten

Der zweite Schritt ist der Aufbau eines Bibliotheksportals mit Nutzersichten. Portale zeichnen sich vor allem durch eine auf bestimmte vordefinierte Nutzergruppen abgestimmte Zugangsstrategie aus. Geht man von Seiten einer Bibliothek davon aus, dass die Suche nach Informationen und Dokumenten fachspezifisch erfolgt, so müssen sowohl herkömmliche Objekte wie Bücher, Zeitschriften, Dias etc. als auch digitale Objekte wie die Digitalisate der anatomischen Sammlung R. Virchows oder elektronische Hochschulpublikationen einer Fachrichtung über ein Portal angeboten werden.

Da es unmöglich ist, für alle Fächer gleichzeitig ein eigenes Portal aufzubauen, wird am Beispiel der Sichtweise eines Kunst- oder Kulturwissenschaftlers eine fachspezifische Recherchemaske definiert. Diese ermöglichst eine gleichzeitige Recherche in den Objektdatenbanken, im Bestand des Dokumentenservers und im Bibliothekskatalog, wobei eben nur die Informationsressourcen, die für einen Kunsthistoriker relevant sind, durchsucht werden sollen. Da die Konfiguration der Nutzersicht, die definiert, welche Zeitschriften, Datenbanken etc. eingebunden werden, sehr vom individuellen Nutzerverhalten abhängt, wird dem Benutzer die Möglichkeit gegeben, seine eigenen Einstellungen in einem Profil zu speichern und für eine spätere Nutzung des Portals zu sichern. Ziel ist es, die Recherchemaske selbst unter Nutzung von XML-Technologien und XML als Datenaustauschformat so flexibel zu gestalten, dass über bestimmte Steuerparameter die eigentliche Sicht (Maske) generisch erzeugt werden kann.

Schaffung eines Zuganges und einer Rechercheschnittstelle

Betrachtet man das Portal, so ist es wichtig, nicht nur die Menge der möglicherweise relevanten Informationen durch Thesauri etc. einzuschränken, sondern auch für die in verschiedenen angesprochenen Datenbanken unterschiedlich benannten Metadatenkategorien Konkordanzen zu finden und zu definieren. Dazu werden gängige Metadatenstandards betrachtet und „übereinander gelegt“. So ist im Bereich der Kunstgeschichte ein Dublin Core Feld „Autor“ in anderen Datenbanken oft der „Künstler“. Es muss sichergestellt werden, dass bei Nutzung einer Recherchemaske für die Suche z. B. zu „Rudolf Virchow“ als „Autor“ sowohl die von ihm geschriebenen Bücher als auch die von ihm angefertigten Präparate („Künstler“, „Präparator“) und Notizbücher (in digitalisierter Form) in den verteilten Datenbanken gefunden werden, nicht jedoch Sekundärliteratur über Rudolf Virchow oder Fotografien, auf denen er zu sehen ist. Diese sollten unter einer inhaltlichen Kategorie auftreten, die natürlich in Dublin Core oder in RAK[8] anders heißt als z. B. bei Nutzung des CIMI-Profils, bei dem man u. U. mehrere Datenbankfelder abfragen muss.

Untersuchung und Darstellung von XML-Standards als Dokumentenformate für Multimediainhalte

Die Kodierung von Dokumenten in XML bildet die Grundlage für eine Langzeitarchivierung elektronischer Dokumente. Die Entwicklungen im XML-

7 http://www.bam-bw.de
8 Regeln für die alphabetische Katalogisierung

Bereich in den letzten Jahren haben unter anderem XML-basierte Beschreibungssprachen wie SVG[9] (Scalable Vector Graphics), SMIL[10] (Synchronized Multimedia Integration Language), MusicML[11] (Music Markup Language), CIMI-Profile sowie X3D (eXtensible 3D) als potentielle neue Normen entstehen lassen.

Daher ist zu überlegen, ob und wie man diese Standards nutzen kann, um zum Beispiel Abbildungen und 3D-Animationen von Sammlungsobjekten mit diesen XML-Formaten zu erfassen. Ein besonderes Augenmerk wird auf die Verwendung des SMIL-Standards gelegt werden, der es ermöglicht, zeitliche Synchronisationen von Medienobjekten vorzunehmen. So könnten zum Beispiel die in Bilddatenbanken gehaltenen Bilder der Kunstgeschichte über eine WWW-Schnittstelle medienneutral für eine Lehrveranstaltung der Kunstgeschichte im SMIL-Format aufbereitet werden.

Sicherung von digitalen Dokumenten

Digitale Dokumente können einfacher als herkömmliche Bilder und Fotografien kopiert und verändert werden. Daher ist es wichtig, auch auf digitale Bilder und Fotografien Methoden anzuwenden, die einerseits verhindern, dass Bilder verfälscht werden, andererseits aber auch klar herausstellen, wer der Urheber der Fotografie oder des Bildes bzw. Videos ist. Dazu sollen digitale Wasserzeichen zur Anwendung kommen. Ähnlich wie in dem bereits erprobten Geschäftsgang „Sicherung der digitalen Dissertationen durch digitale Signaturen und Zeitstempel“, soll die Nutzung von Verfahren zum Anbringen und zur Überprüfung digitaler Wasserzeichen für die Humboldt-Universität standardisiert werden. Dabei kann auf die am Institut für Integrierte Publikations- und Informationssysteme der GMD Darmstadt (GMD-IPSI) gemachten Erfahrungen mit Signaturen für Bilder und Videodaten zurückgegriffen werden.

Erarbeitung von Vorgaben, Empfehlungen und Richtlinien für Autoren

Der bereits an der Humboldt-Universität etablierte Workflow zur Erstellung und Konvertierung von XML aus Formaten anderer proprietärer Textverarbeitungssysteme ist bisher sehr gut für textorientierte Dokumente anwendbar. XML-Objekte können aus herkömmlichen Multimediaobjekten erzeugt werden, zum Beispiel zur Abspeicherung von Vektorgrafiken mit der SVG.DTD (Scalable Vector Graphics), von 3D-Visualisierungen mit der X3D.DTD und von mathematischen Formeln mit der MathML.DTD etc. Dazu müssen jedoch Anleitungen erarbeitet werden und Schulungen im Umgang mit diesen Methoden und Software-Werkzeugen durchgeführt werden, damit alle Autoren multimedialer Dokumente diese auch nutzen können.

Selbstverständlich müssen diese Anleitungen in das bisher bestehende Schulungskonzept integriert werden. Bisher besteht es aus Lehrgängen für digitale Dissertationen und aus Kursen, in denen in die Gestaltung verschiedenartiger Medien eingeführt wird.

9 siehe Definition des W3C: http://www.w3.org/Graphics/SVG/
10 siehe Definition des W3C: http://www.w3.org/AudioVideo/#SMIL
11 siehe http://195.108.47.160/3.0/musicml/index.html

Literatur

[1] Martin, N.; Schirmbacher, P.: Die elektronische Publikation von Dissertationen an der Humboldt-Universität zu Berlin. *RZ-Mitteilungen*. 1997, Heft 14, http://www.hu-berlin.de/rz/rzmit/rzm14/rzm14_14.html

[2] Dobratz, S.: Die „Humboldt-Dissertationen“ im Internet. *RZ-Mitteilungen*. 1998, Heft 16, http://www.hu-berlin.de/rz/rzmit/rzm16/4.html

[3] Dobratz, S.: XML – eXtensible Markup Language. *RZ-Mitteilungen*. 1999, Heft 18, http://www.hu-berlin.de/rz/rzmit/rzm18/6.html

[4] Dobratz, S.: XML – ein Überblick über das Chaos. *RZ-Mitteilungen*. 2000, Heft 20, http://www.hu-berlin.de/rz/rzmit/rzm20/5.pdf

[5] Müller, U.: Dublin Core: Metadaten für WWW-Inhalte. *RZ-Mitteilungen*. 2000, Heft 20, http://www.hu-berlin.de/rz/rzmit/rzm20/15.pdf

[6] Schulz, Matthias: ProPrint. *RZ-Mitteilungen*. 2001, Heft 21

Susanne Dobratz
dobratz@rz.hu-berlin.de

So war es zu lesen in den RZ-Mitteilungen Heft Nr. 10/1995

War man bei den gedruckten Ausgaben, die diesen Volltextdatenbanken zugrunde liegen, noch auf Register oder mühsames Durchblättern des gesamten Textes angewiesen, so kann man nun in Sekundenschnelle feststellen, in welchen Werken und in welchem Kontext etwa Goethe von Drachen spricht. Das Bundeskartellamt betrachtet deswegen CD-ROM-Datenbanken als ein qualitativ so stark von der gedruckten Ausgabe abweichendes Erzeugnis, daß es diese von der für Verlagserzeugnisse möglichen Preisbindung ausnehmen möchte.

So war es zu lesen in den RZ-Mitteilungen Heft Nr. 18/1999

Das Publizieren von Datenbanken wird sich zum besonders gut genutzten Anwendungsgebiet von XML entwickeln, da die Unterstützung objektorientierter Konzepte wie die Darstellung der inneren Struktur von Objekten oder deren hierarchischer und Linkbeziehungen nun einfacher möglich ist.