

Document Type Definitions zur Erschließung von Gattungen des Barock im Internet. Ein Projekt an der Herzog August Bibliothek Wolfenbüttel¹

Seit Januar 2001 arbeitet die Herzog August Bibliothek an dem mit Mitteln der *Deutschen Forschungsgemeinschaft (DFG)* geförderten Projekt *Barock DTDs – Document Type Definitions zur Erschließung barocktypischer Gattungen im Internet*.

Mit Hilfe von Fachwissenschaftlern sollen fünf für das 17. Jahrhundert charakteristische Textgattungen mit Blick auf ihre gattungstypischen Eigenschaften bearbeitet und prototypische Document Type Definitions (DTDs) für die Erstellung von XML-Dokumenten entwickelt werden.

1. Historische Drucke im Internet

Anspruch und Grenzen der retrospektiven Digitalisierung von Druckwerken und ihrer computergestützten Editionen sind in den letzten Jahren mit wachsendem Interesse diskutiert worden. Angesichts der vielen isolierten Digitalisierungsprojekte² - auch im Bereich des alten Buches - stellt sich die Frage nach einheitlichen Erschließungsstandards. Imagesequenzen allein reichen nicht aus, um einen Druck im Internet verfügbar zu machen. Die Diskussion dreht sich um die geeignete Form der Erschließung, um Metadaten, um die Frage, wie globale oder spezielle Suchmaschinen Dokumente im Netz zu finden vermögen, wie Daten über digitalisierte alte Drucke optimal verwaltet, ausgetauscht oder auch präsentiert werden können.

Konsens besteht darüber, dass ein besonderer Gewinn bei elektronischen Publikationen in der Erschließung liegt. Über die Bereitstellung der reinen Imagesequenz hinaus soll dem Benutzer das elektronische Buch inhaltlich geöffnet werden, um im neuen Medium einen Mehrwert gegenüber der gedruckten Form zu erzielen.

In dem von der DFG geförderten Projekt *Barock DTDs – Document Type Definitions zur Erschließung barocktypischer Gattungen im Internet* will die Herzog August Bibliothek daher, über die bibliothekarische Formal- und Sacherschließung, die in jedem Online-Katalog zu finden ist, hinaus projektbezogen weitergehende Recherchemöglichkeiten anbieten. Für fünf Textgattungen des Barock soll eine terminologische und strukturelle Grundlage geschaffen werden für die zukünftige netzbasierte digitale Publikation von Drucken des 17. Jahrhunderts.

Voraussetzung für die Präsentation von alten Drucken im Internet ist zunächst einmal ihre buchschonende Digitalisierung. Wir digitalisieren die Drucke in Farbe. So wird über den Text hinaus mit der möglichst originalgetreuen Farbkopie auch ein angemessener Eindruck von der Materialität des Druckes vermittelt.

Im Netz zu sehen ist eine Imagesequenz der einzelnen Buchseiten des Originaldrucks. Wir stellen also ein elektronisches Faksimile des Buches im Internet bereit, das mit Hilfe eines Dokumentenmanagementsystem zum Blättern geöffnet und Seite für Seite durchgeblättert werden kann.

2. Verwendung von Standards

Auf der Grundlage der Auszeichnungssprache *Standard Generalized Markup Language (SGML)* und davon abgeleitet *eXtensible Markup Language (XML)* sind internationale

¹ Die nachfolgende Darstellung basiert auf einem Vortrag, der am 22.02.02 im Rahmen der Internationalen Arbeitstagung der Arbeitsgemeinschaft für Germanistische Edition, der Arbeitsgemeinschaft philosophischer Editionen und der Fachgruppe Freie Forschungsinstitute in der Gesellschaft für Musikforschung *Autor, Autorisation, Authentizität* in Aachen gehalten wurde. Hier wird eine gekürzte Fassung wiedergegeben. Der vollständige Text wird im *Jahrbuch für Computerphilologie 2003* erscheinen. (<http://computerphilologie.uni-muenchen.de/ejournal.html>)

² Eine Übersicht der DFG-geförderten Digitalisierungsprojekte ist unter folgender Adresse zu finden: http://www.sub.uni-goettingen.de/gdz/de/projects/vdf_verzdokuserv_1.html

Standards und Quasistandards zur Strukturierung, Beschreibung und Erschließung von elektronischen Dokumenten entstanden, die wir nutzen.

In unserem Projekt arbeiten wir mit der vom *World Wide Web Consortium (W3C)* empfohlenen Standardauszeichnungssprache XML.³ Mit der Festlegung auf XML ist zunächst die formale Grundlage für eine netzbasierte Interoperabilität gewonnen. Die Standardisierung von Dokumentenstrukturen, die eine einheitliche Recherche und Präsentation inhaltlicher Aspekte erlaubt, setzt neben der Verständigung auf eine formale Auszeichnungssprache die Entwicklung von Regelwerken für diese Sprache, sprich fach- bzw. gattungstypische DTDs (Document Type Definitions) voraus. DTDs beschreiben in formaler Notation die logische Struktur eines bestimmten Dokumententyps, sind also das Regelwerk für XML-Strukturen. Erst mit solchen DTDs läßt sich die Homogenität sicherstellen, die ein einheitliches Retrieval und ein netzbasiertes Arbeiten ermöglichen. DTDs ermöglichen das automatische Validieren von Dokumenten, die Bildung von Standards, einheitliche Recherche und die Nutzung auch dezentral erfasster Daten über das Internet.

Für den Bereich der Geisteswissenschaften ist die *Text Encoding Initiative (TEI)*,⁴ besonders hervorzuheben in dem Bemühungen, eine Syntax zur Repräsentation geisteswissenschaftlicher Texte, differenziert nach literaturwissenschaftlichen Textsorten, zu konzipieren. Ursprünglich auf SGML aufbauend, bietet die *TEI* jetzt auch ein XML-Vokabular für das Auszeichnen von Texten in den Geisteswissenschaften an. Obwohl die *TEI* in erster Linie darauf abzielt Text auszuzeichnen, erlauben die *TEI*-Guidelines auch das Beschreiben und Indexieren von Images. Inzwischen hat sich *TEI* auf internationaler Ebene in vielen Projekten⁵ als Standard etabliert, da sie sowohl plattform- wie medienunabhängigen Zugriff auf elektronisch bereitgestellte Dokumente ermöglicht.

Für spezifische Textsorten des 17. Jahrhunderts stehen aber bislang keine DTDs zur Verfügung. In unserem Projekt wollen wir daher in Zusammenarbeit mit Fachwissenschaftlern fünf für das 17. Jahrhundert charakteristische Textgattungen mit Blick auf ihre gattungstypischen Eigenschaften bearbeiten und prototypische Document Type Definitions auf Grundlage der *TEI*-DTD erstellen. Ein großer Vorteil der *TEI*-DTD besteht in ihrer Offenheit, die eine flexible Erzeugung bzw. Wiedergabe von strukturierten Inhalten ermöglicht. Durch Spezifikation der Verwendung einzelner Elemente und die Festlegung bestimmter Attributwerte der *TEI*-DTD haben wir Strukturvorgaben nach den Anforderungen der von uns bearbeiteten Gattungen entwickelt. Damit soll die strukturelle Eigenart der Quelle berücksichtigt werden und zugleich der Vorteil des internationalen Standards genutzt werden.

Ausgewählt wurden hierzu die Gattungen *Illustrierte Flugblätter, Emblembücher, Kalender und Prognostiken, Gebet- und Gesangbücher* und *Pest- und Seuchenschriften*. Diese Gattungen wurden zum einen deshalb ausgewählt, weil es sich um von der Frühneuzeitforschung besonders nachgefragte Literatur handelt. Zum anderen sollten die einzelnen Gattungen relativ einheitlich faßbar und beschreibbar sein und unter den Neuerwerbungen der Herzog August Bibliothek repräsentativ vertreten sein, d.h. in gewisser Quantität, aber auch mit einzelnen herausragenden Stücken, die für die Forschung relevant sind. Wir hoffen mit der Entwicklung standardisierter Schnittstellen auf der Basis akzeptierter formaler Standards wie XML und *TEI* wichtige Impulse für eine netzbasierte Erschließung dieser Textgattungen zu geben. So könnten mit Hilfe dieser DTDs verteilt erstellte XML-Daten sowohl unmittelbar als auch in einer in HTML konvertierten und mit Meta-Tags versehenen Form Grundlage für Internet Suchmaschinen werden.

³ <http://www.w3c.org/XML/>

⁴ <http://www.tei-c.org/>

⁵ Als ein Beispiel sei hier das *Emblem Project Utrecht* genannt: <http://www2.let.uu.nl/emblems/html/index.html>.

3. Barock-DTDs – Das Arbeitskonzept

Wie stellt sich die Erschließungsarbeit nun konkret dar? Am Beispiel von zwei der fünf Gattungen soll das Arbeitskonzept erläutert werden. Die illustrierten Flugblätter und die Emblembücher mit ihren komplexen Text-Bildbeziehungen scheinen hierzu besonders geeignet. Beide geben dem kulturhistorisch Interessierten und dem Fachwissenschaftler vielfältige Zugänge zum Verständnis der Epoche. Die Erschließung und einfache Zugänglichkeit dieser Quellen ist die Voraussetzung für ihre intensive Nutzung.

Nach der Digitalisierung der Drucke werden die Digital-Master zur Präsentation im Internet von ca. 20 MB auf rund 150 KB als JPG komprimiert. Die Zugänglichkeit der Drucke im Internet wird von der Herzog August Bibliothek garantiert, ebenso die beständigen URLs. Damit wird die Zitierbarkeit der elektronischen Gesamtdokumente und der einzelnen Images sicher gestellt. Eine wichtige Voraussetzung, um mit der elektronischen Publikation arbeiten zu können .

Wenn die Images vorliegen, erfolgt die Inhaltserschließung anhand der zuvor erstellten gattungsspezifischen DTD. Dabei setzt sich jede DTD aus drei Bereichen zusammen:

1. Metadaten
2. Gattungsspezifische Bestandteile
3. Inhaltliche Erschließung

1. Metadaten:

Für das Retrieval und die Identifikation von Quellen im Netz sowie für deren langfristige Archivierung sind Metadaten⁶ nach internationalen Standards von besonderer Bedeutung. Auf der Grundlage der Richtlinien der *TEI* wurden im Projekt die bibliographischen Metadaten-Elemente für die elektronischen Dokumente ausgewählt. Sie berücksichtigen die Katalogisierungsdaten im Bibliothekssystem PICA⁷, die im Onlinekatalog benutzt werden. Diese Metadaten sind hierarchisch strukturierter Inhalt des *TEI*-Elements *TEI-Header*. Ein eigens hierfür entwickeltes Script konvertiert die relevanten Katalogisierungsdaten und fügt sie ein in *tei*-konforme XML-Dokumente, die es zugleich nach der DTD erzeugt. Eine Ergänzung um *Dublin Core* Elemente⁸ wird bei der Transformation nach HTML vorgenommen.

2. Gattungsspezifische DTD-Bestandteile:

Die Metadaten zum elektronischen Dokument haben unabhängig von der jeweiligen Gattung der Quelle eine einheitliche Struktur in der DTD. Die Bestandteile zur Erfassung und zur inhaltlichen Erschließung der zugrundeliegenden Quelle differieren dagegen nach deren Gattungszugehörigkeit. Das bedeutet, dass zuerst für jede Gattung eine grundsätzliche Dokumentenanalyse durchgeführt werden muss, in der die Gattungsmerkmale definiert und zu DTD-Bestandteilen formalisiert werden.

3. Inhaltliche Erschließung:

Zur inhaltlichen Erschließung benutzen wir im wesentlichen das *TEI*-Element *<index>*. Dieses Element kann flexibel auch zur Markierung von Bereichen benutzt werden, die hierarchische Kapselungen überlagern. Diese Eigenschaft ist innerhalb von Dokumenten mit seiten- bzw. bildorientierter Struktur von großer Bedeutung. Gattungsspezifischen

⁶ Metadaten enthalten Anweisungen für Web-Server, Web-Browser und Suchprogramme im Internet, sie können Angaben zum Verfasser oder zum Inhalt der Datei enthalten

⁷ <http://www.gbv.de/>

⁸ Dublin Core ist ein von einer internationalen Expertengruppe definiertes Elemente-System für Metaangaben: <http://dublincore.org/>

Erfordernissen trägt es durch seine Einbettung in entsprechend qualifizierten Elementen und durch Attribute Rechnung. Letztere ermöglichen auch eine Zuordnung zu unterschiedlichen Indices, die vielfältig spezialisierte Retrievalfunktionen zulassen.

4. Erschließung der Flugblätter

Am Beispiel der Flugblätter sollen die einzelnen Arbeitsschritte kurz erläutert werden. Zunächst erfolgt die Dokumentenanalyse. Das illustrierte Flugblatt besteht in der Regel aus drei Teilen: Überschrift, Bildsegment und Textsegment. Diese drei Bestandteile spiegeln sich in der DTD-Struktur wieder, bilden das gattungsspezifische Gerüst.

Für jedes Dokument werden in den TEI-Header die bibliographischen Daten aus den vorhandenen Katalogaufnahmen des Online-Kataloges übernommen und die technischen Angaben zum elektronischen Dokument hinzugefügt. Damit sind Angaben wie Erscheinungsort und -jahr, Autor, Titel etc. zur Recherche verfügbar. Dieser Bereich der Metadaten ist für alle Gattungen gleich.

Zu diesen Angaben fügen wir eine systematische Einteilung in Anlehnung an die zeitgenössische Fächerhierarchie *Theologica, Ethica, Politica, Physica, Casualia* hinzu.⁹ Wenn inhaltlich mehrere Bereiche berührt sind, kann die Zuordnung auch zu mehreren systematischen Gruppen erfolgen.

Das Flugblatt wird über Schlagwörter erschlossen, die die Gesamttendenz des Blattes beschreiben sollen, z. B. Dreißigjähriger Krieg. Die Beschreibung des Bildteils erfolgt durch die Vergabe von Bildschlagwörtern (Beutel, Kugel usw.). Sofern im Bild auch Text vorhanden ist, wird dieser in Bildstichwörtern aufgenommen. Die Erfassung der Bildelemente mittels *Iconclass*¹⁰ ist vorgesehen, muss aber nicht ausgefüllt sein.

Im dritten Arbeitsschritt erfolgt die Erschließung des Textsegmentes. Bei der Texterschließung werden die Leitbegriffe in heutiger Orthographie (level1) und in der originären Schreibweise (level2) als Stichwörter aufgenommen. Daneben bietet die DTD jedoch auch die Möglichkeit Volltext (z. B. ausgewählte Textteile) einzugeben.

Die Dokumentenerschließung erfolgt wie erläutert in XML. Dabei werden die XML-Editoren *Xmetal* und *XML Spy* eingesetzt.

Ist die Dokumentenerschließung abgeschlossen und als XML-Struktur nach der DTD erfasst, ist ein weiterer Arbeitsschritt nötig, um die Ergebnisse auch dem Internet-Benutzer zugänglich zu machen.

Die XML-Strukturdaten müssen bearbeitet werden, um das XML-Dokument in eine Webseite, in ein HTML-Dokument zu transformieren. Dies geschieht mittels XSLT, *eXtensible Stylesheet Language-Transformations* (auch ein XML-Format), das es ermöglicht, entweder ein neues XML-Dokument oder ein HTML-Dokument für die Internetpräsentation zu erstellen. Auf diesem Weg wird das Dokument mit *Dublin Core*-Metadaten angereichert und kann für die Recherche (Indexierung XML-Instanzen) in eine MySQL-Datenbank eingebracht werden.

5. Die Erschließung der Emblembücher

⁹ Vgl. hierzu Wolfgang Harms (Hrsg.): Deutsche illustrierte Flugblätter des 16. und 17. Jahrhunderts. Tübingen: Niemeyer, 1980 -.

¹⁰ Iconclass ist ein Klassifikationssystem mit ca. 28 000 Definitionen und einem alphabetischen Index zur Bilderschließung, das vielfach angewendet wird: <http://iconclass.bureau.knaw.nl/>

Die Arbeitsschritte für die Emblembücher sehen entsprechend aus.

Auch hier steht an erster Stelle die Dokumentenanalyse: Die Embleme bestehen in der Regel aus drei Teilen: dem themaandeutenden *Motto*, der gegenstandsdarstellenden *Pictura* und der auslegenden *Subscriptio*, die auf einer Buchseite zu finden sind. Hinzu kommen die Erläuterungen oder anderer Text zwischen den Emblemseiten. Nicht jedes Emblembuch folgt jedoch dieser Einteilung. Es gibt ebenso Emblembücher mit mehreren *Picturae* auf einer Buchseite. *Motto* und die *Subscriptio* zu den einzelnen Emblemen sind auf andere Seiten verteilt. Auch diese Fälle müssen in der DTD berücksichtigt werden, damit bei der Datenbankrecherche und in der HTML-Darstellung korrekte Ergebnisse erzielt werden können. Die Abweichung vom dreigliedrigen Idealtypus stellt für die DTD-Spezifikation eine Herausforderung dar. Nach der Analyse zahlreicher Emblembücher haben wir uns entschieden, die Strukturierung der XML-Dokumente in erster Linie an der Form des jeweiligen Buches nicht an der des Einzelemblems auszurichten. Die Zugehörigkeit der einzelnen Teile werden bei getrennter Verteilung durch verweisende Attributwerte gekennzeichnet. Man soll die Embleme in ihrem originären Kontext wie auch isoliert betrachten können.

Bei der Bearbeitung der Emblembücher haben wir verschiedene Erschließungstiefen vorgesehen und ausprobiert. Abhängig von den beabsichtigten Arbeitsergebnissen, von den zur Verfügung stehenden finanziellen und personellen Ressourcen können so unterschiedliche Erfassungs- und Erschließungsstrategien angewendet werden.

6. Ausblick

Mit dem Ziel, die terminologische und strukturelle Basis für eine netzbasierte digitale Publikation von Drucken aus dem 17. Jahrhundert zu schaffen, ist unter Berücksichtigung internationaler Standards wie XML und *TEI* ein Konzept entstanden, das es erlaubt, im Internet eine aus Images bestehende Faksimile-Edition auf der Grundlage der eher Text- bzw. semantisch orientierten *TEI*-DTD zu erstellen. Dabei bildet die DTD die unmittelbare Voraussetzung dafür, die Konsistenz nicht nur lokal, sondern auch dezentral erfasster Daten mittels Validierung zu gewährleisten und sie in einem zentralen Datenbanksystem zusammenfassen zu können.

Dem Nutzer sollen möglichst vielfältige Zugangswege zum Dokument eröffnet werden. So sind die digitalisierten Drucke nicht nur über das Internet (Suchmaschinen) und über die Datenbank erreichbar, sondern auch über den OPAC der Bibliothek und die VD17-Datenbank. In einer Signaturen-Linkliste¹¹ werden alle im Projekt digitalisierten Titel zusammengeführt. Hier kann man sich einen Überblick verschaffen, welche Titel bereits elektronisch vorliegen.

Zum Abschluss sei ein Blick auf die erreichten Arbeitsergebnisse gestattet.

Bisher haben wir 200 Drucke mit ca. 30 000 Images digitalisiert. Wir haben ein Konzept für die Verbindung von Text und Images entwickelt. Ein PICA-*TEI*-Header Konverter und eine XML-Datenbankschnittstelle sind programmiert worden. Es wurden XSLT-Transformationsscripts für die einzelnen Gattungen geschrieben.

Wir haben für 30 Flugblätter die Dokumentenanalyse und die Erschließung fertiggestellt. Die Dokumentenanalyse für die Emblembücher, für Kalender und Prognostiken sowie für die Pest- und Seuchenschriften und die Gebet- und Gesangbücher ist abgeschlossen.

Erschlossen wurden bisher zehn Emblembücher, zwölf Kalender, zehn Pest- und Seuchenschriften und vier Gebet- und Gesangbücher.

Die Entwicklung der Datenbank ist soweit fortgeschritten, dass wir sie im Februar 2003 präsentieren können.¹²

¹¹ <http://www.hab.de/forschung/de/barock-dtd/SigDTD.html>

¹² Der aktuelle Arbeitsstand ist auf unserer Internet Projektseite zu ersehen:
<http://www.hab.de/forschung/de/barock-dtd/index.htm>

Unser Ziel ist es, prototypische DTDs für einzelne barockspezifische Gattungen auszuformulieren. Für die fünf Gattungen Flugblatt, Emblembuch, Kalender, Gebetbuch und Seuchenschrift werden wir am Ende des Projekts eine solche prototypische Ausformulierung vorlegen können. Mit der Verwendung von Standards wie *TEI* und XML sind die Grundlagen für Interoperabilität geschaffen. So soll es möglich sein, dass zukünftig *TEI* basierte inhaltlich gleichartige digitale Sammlungen auch zusammengeschlossen werden können, nicht nur auf bibliographischer sondern auch auf inhaltlicher Ebene.

Andrea Opitz (Wolfenbüttel)