Erfahrungen mit internationalen Standards bei der Handschriftenbeschreibung: Die Verwendung von MASTER bei den CEEC

von PATRICK SAHLE

Vorbemerkungen / Einleitung

Die Verwendung eines Standards zur Handschriftenbeschreibung innerhalb des Projektes Codices Electronici Ecclesiae Coloniensis (CEEC) muß sich nach den Grundzielen des Vorhabens richten: zentrale Aufgabe ist es hier, eine Erschließung des Manuskripten-Bestandes auf der Basis bereits verfügbaren Fachwissens durchzuführen und *nicht*, eigenständige handschriftenkundliche Forschungen durchzuführen. Durch eine deutlich erhöhte Verfügbarkeit des vorhandenen Wissens und durch gänzlich andere Nutzungsformen soll dann die weitere Forschung und damit die Gewinnung neuer Erkenntnisse gefördert werden.

Die zu verarbeitenden Informationen über die einzelnen Handschriften ergeben sich aus den gedruckt, maschinenschriftlich, handschriftlich oder elektronisch vorliegenden Katalogen, sowie Teilen der Sekundärliteratur. Für die Verwaltung von Informationen bildet in der Praxis das relationale Datenmodell derzeit noch das am weitesten durchgesetzte Paradigma.[1] Eine Übernahme der Informationen in eine relationale Datenbank mit vorgegebenen Feldern und exakten Definitionen der erlaubten Datentypen wäre im konkreten Fall der Vielfalt und Heterogenität der vorhandenen - und zunächst als autoritativ anzunehmenden - Informationen jedoch nicht gerecht geworden. Die Umformung und Aufgliederung von Fließtexten (wie sie in Katalogen zuweilen vorkommen) als wenig formaler Datenstruktur in typisierte Einzelinformationen hätte durch die Ent(kon)textualisierung und die Konzeptdivergenz zwischen Katalog und Datenbank nicht nur Informationsverluste mit sich gebracht. Sie wäre auch wegen der permanent nötigen wissenschaftlichen Begleitung der Umarbeitung zu zeitaufwendig gewesen.

Eine Digitalisierung der Kataloge als reine Volltexte, oder sogar nur als Faksimiles, hätte aber nur einen beschränkten Gewinn an Verfügbarkeit und fast keinen Fortschritt in den Nutzungsformen bringen können. Gesucht war deshalb eine Datenstruktur, die einerseits einfach, flexibel und textnah sein sollte, um

[1] Diesem Modell folgt z.B. auch die "Handschriftendatenbank", die gemeinsam von Bayerischer Staatsbibliothek, Staatsbibliothek Berlin und dem Bildarchiv Foto Marburg aufgebaut wird. Siehe die Einstiegsseite <http://www.fotomr.uni-marburg.de/hs-bank.htm>.

tenstruktur, die einerseits einfach, flexibel und textnah sein sollte, um heterogene (Text-)Daten verwalten zu können, und die andererseits ein gewisses Maß an Strukturierung und Formalisierung erlauben sollte, um darauf komplexe Zugriffs- und Nutzungsformen aufbauen zu können.

Aus den genannten Gründen fiel die Wahl im Projekt CEEC auf einen Datenstandard mit dem etwas weit ausgreifenden Akronym MASTER. Im folgenden wird kurz zusammengefaßt werden, (1.) was MASTER ist und (2.) wie dieser Standard innerhalb von CEEC verwendet wird.

Der MASTER-Standard

[MASTER-Standard:] Grundlagen und Ziele

"MASTER"[2] steht für "Manuscript Access through Standards for Electronic Records" und ist ein EU-finanziertes Projekt mit einer Laufzeit von zweieinhalb Jahren (Januar 1999 bis Juni 2001). Es wird getragen von einigen zentralen Projektpartnern[3] und einem "Kranz" assoziierter Partner, jeweils aus dem Bereich der europäischen Bibliotheken und Universitäten (dabei vor allem sogenannte 'Humanities-Computing'-Abteilungen oder Text- und Quellen-orientierte Spezialinstitute mit technologischer Ausrichtung).

Die Ermöglichung eines "allgemeinen" Handschriftenkataloges ist das umfassende Hauptziel des Projektes.[4] Um dies zu erreichen, wird vor allem die Definition eines Standards für maschinenlesbare Handschriftenbeschreibungen ausgearbeitet. Darüber hinaus soll vom Projekt Software zur Erstellung und Kontrolle (Validierung) dieser Katalogisate bereitgestellt, Tests mit Beispielkatalogisaten durchgeführt und diese selbst im Internet angeboten werden.

Technische Grundlage sind die sogenannten Auszeichnungs- oder Markup-Sprachen, genauer gesagt SGML und sein Derivat XML.[5] Diese Wahl wird begrün-

[2] Die umfangreiche Projektdokumentation mit allen Ressourcen, die von MASTER bereitgestellt werden findet sich unter <http://www.cta.dmu.ac.uk/projects/master/>.

[3] Koordinierend: Centre for Technology and the Arts, De Montfort Universität, Leicester. Weitere zentrale Partner: Koninklijke Bibliotheek, Den Haag; L'Institut de recherche et d'histoire des textes, Paris/Orleans; Arnamagnaean Institute, Copenhagen; Humanities Computing Unit, Universität Oxford.

[4] Der Begriff "allgemein" meint hier tatsächlich - zumindest theoretisch - "allumfassend" - siehe dazu die Bemerkungen weiter unten.

[5] SGML (Standard Generalized Markup Language) ist ein seit 1986 veröffentlichter internationaler Standard (ISO 8879) zur Beschreibung von Dokumentstrukturen. XML ist eine vom World-Wide-Web-Consortium betreute vereinfachte Fassung von SGML, welche die Entwicklung beliebiger spezieller Auszeichnungs-

det mit dem Hinweis darauf, daß es sich bei SGML um einen international akzeptierten und verbreiteten Standard handelt, der es ermöglicht, einfache wie komplexe Beschreibungen zu integrieren, und der zugleich feldorientierte und textorientierte Daten zuläßt. Die Auszeichnungssprachen haben ihre Stärke in der Verwaltung textnaher Daten. Innerhalb des Projektes gab es am Anfang allerdings auch Überlegungen, als Eingabeprogramm mit Microsofts ACCESS eine relationale Datenbank zu verwenden. SGML/XML-Datenmodelle bilden grundsätzlich hierarchische Strukturen ab, die in "Elemente" zerfallen, die selbst wieder Elemente enthalten können. Eine solche Struktur wird oft als "Elementbaum" charakterisiert, mit einem Wurzelelement und fortlaufenden Verästelungen. Die Nähe einer solchen Struktur zu fortlaufenden Texten wird deutlich, wenn gesagt werden kann, daß die Datenstruktur *über* einen Text (hier z.B.: eine Handschriftenbeschreibung) gelegt wird. Der Text wird zwar (technisch gesehen) in Elemente zerlegt, er wird aber nicht aufgelöst oder desintegriert: die Elemente bilden nach wie vor einen fortlaufenden Text. Die Beschreibung der (Text-)Daten, d.h. die Umformung des Textes in eine formale Datenstruktur verläuft dabei von allgemeinen Gliederungsebenen zu immer detaillierteren Einzelheiten und kann in diesem Prozeß auf jeder beliebigen Stufe stehen bleiben. Das bedeutet auch, daß es keinen Zwang zu einer vollständigen oder auch nur gleichmäßigen Anwendung oder Übertragung des Datenmodells auf die vorhandenen Informationen gibt.

Mit der Wahl von SGML/XML verfolgt das MASTER-Projekt das Ziel einer möglichst allgemeinen Verwendbarkeit des Standards, der sowohl softwareunabhängig, als auch indifferent gegenüber den späteren Nutzungsformen ist. Einen hohen Stellenwert hat deshalb auch die möglichst hohe Kompatibilität zu anderen Standards. Hier ist vor allem die Text Encoding Initiative (TEI) zu nennen, innerhalb der es ebenfalls eine Arbeitsgruppe für die Entwicklung einer formalen Definition für den Dokumententyp "Handschriftenbeschreibung" gibt.[6] Durch eine enge Kooperation von MASTER und TEI mit Abgleich der Positionen und Zwischenergebnisse ist sichergestellt, daß am Ende ein identisches Ergebnis von beiden Seiten verabschiedet werden wird.

---

sprachen für das Internet ermöglichen soll. Zum aktuellen Stand der XML-Entwicklung siehe <http://www.w3.org/XML/>.

[6] Die Text Encoding Initiative (TEI) wird inzwischen von einem Konsortium, angesiedelt an vier Universitäten (Oxford, Bergen, Virginia, Brown-University), geleitet; siehe <http://www.tei-c.org/>. Informationen über die Fachgruppe für Handschriftenbeschreibungen innerhalb der TEI sind nicht leicht zu finden, einen Überblick über die Zusammenarbeit mit MASTER gibt das letztere Projekt: <http://www.cta.dmu.ac.uk/projects/master/teimaster.html>.

Inhaltlich betrachtet zielt der MASTER-Standard nicht auf die Entwicklung einer neuen Katalogisierungsrichtlinie für Handschriften, sondern versucht die Integration verschiedener bestehender Konzepte unter einer verallgemeinernden Perspektive. Die Berücksichtigung der unterschiedlichen nationalen und fachlichen Traditionen führt dabei zwangsläufig zu einer gewissen "Weichheit" bzw. "Unbestimmtheit" der Definition: weniges ist strikt vorgegeben, oft sind für gleichartige Informationen verschiedene formale Ausprägungen möglich. Diese Weichheit und Allgemeingültigkeit der Definition zeigt sich auch bei einem vergleichenden Blick in die "Richtlinien Handschriftenkatalogisierung" der DFG:[7] Die dort gemachten Vorschläge kollidieren in keinster Weise mit MASTER, sie könnten vielmehr als - vergleichsweise sehr einfaches - Subset innerhalb von MASTER beschrieben und formalisiert werden. In Einzelfällen wären aber andererseits - dort wo die DFG-Richtlinien detailliertere Muster vorgeben - kleinere Ergänzungen an den MASTER-Vorgaben vorzunehmen.[8]

[MASTER-Standard:] Verwendungszwecke

MASTER zielt auf zwei durchaus verschiedene Vorgehensweisen bzw. Arbeitsprozesse zur Erstellung elektronischer Handschriftenkataloge. Einerseits läßt sich - u.U. als konzeptionelle Auswahl aus dem Gesamtstandard - ein Eingabeformular ableiten, um (Neu-) Katalogisierungen zu bisher unerschlossenen Handschriften durchzuführen. Andererseits bietet es ein Modell, in das alte, gedruckt vorliegende Katalogisate durch Retrokonversion überführt werden können.

Das bereits genannte Fernziel des *einen allgemeinen* Handschriftenkataloges verbindet diese beiden Stoßrichtungen, impliziert aber kein unkontrollierbares neues Großunternehmen mit ewiger Laufzeit. Die Integration potentiell aller Handschriftenkatalogisate ist eine offene Option; sie erzwingt keine zentrale Leitung und benötigt keinen großen organisatorischen Rahmen. Da der Standard nah an bestehenden Traditionen und der traditionellen Form "Fließtext" bleibt, sind keine speziellen Verwendungsformen fest vorgegeben. Nach MASTER erstellte oder ausgezeichnete Handschriftenbeschreibungen lassen sich in verschiedene Nutzungszu-

[7] Herausgegeben vom DFG-Unterausschuß für Handschriftenkatalogiserung. Z.B. Bonn-Bad Godesberg [5]1992.

[8] Dies könnte z.B. bei der genaueren Scheidung von 'Incipit', 'Initium' und 'Rubrizierung' gemacht werden. Nützlich wäre eventuell auch eine Einführung detaillierterer Typologien in bereits vorhandenen MASTER-Elementen, wie sie bei den DFG-Richtlinien für den Bereich 'Buchschmuck' genannt werden.

sammenhänge einbinden und bei Bedarf sogar in traditioneller Form als Katalog (einschließlich der üblichen Register) ausdrucken.

Selbst ein Zwang zur sklavischen Beachtung der Vorgaben des Standards besteht nicht. Vielmehr können, speziellen Traditionen oder situativen Vorgaben folgend, auch dezentrale Handschriftenkataloge an Einzelinstitutionen oder Verbünden - gegebenenfalls in teilweiser Abweichung von den einzelnen MASTER-Definitionen - erstellt werden - genau dies wird z.B. in CEEC umgesetzt. Durch die Abweichung vom Standard durch Verengung oder Erweiterung einzelner Bereiche wird dabei die Möglichkeit der späteren Integration verschiedener Kataloge in ein Metasystem *nicht* aufgehoben.

[MASTER-Standard:] die technische Perspektive

Der MASTER-Standard ist eine Document Type Definition (DTD) für SGML/XML-Dokumente. Eine DTD beschreibt die Struktur eines Dokument-Typs, hier also, welche Elemente in einem Handschriftenkatalogisat vorkommen können und wie diese Elemente möglicherweise hierarchisch geordnet oder ineinander enthalten sind. Die formale Definition beschreibt außerdem zusätzliche Attribute zu einzelnen Elementen, die eine weitere Spezifikation der Informationen erlauben.[9] Eine DTD bildet ein Datenmodell ab und gibt somit einen formalen Rahmen für ein inhaltliches Konzept; sie ermöglicht es aber andererseits auch, elektronische Katalogisate auf ihre Gültigkeit, das ist ihre Konformität zum Standard, zu prüfen.

Ein wesentliches - nicht markup-spezifisches - Problem ergibt sich aus dem Umstand, daß die in der DTD formal definierten Elemente außerhalb der DTD zu ihrer korrekten Verwendung sprachlich erläutert werden müssen. Angesichts der Tatsache, daß in MASTER möglichst vielen nationalen und fachlichen Traditionen der Katalogisierung Rechnung getragen werden soll, verstärkt sich die 'Weichheit' der Definitionen noch um eine gewisse Mehrdeutigkeit der Interpretation einzelner Elemente und ihrer jeweiligen Verwendung.

Interpretationsspielraum und Interpretationsdivergenzen zeigten sich - um kurz der Praxis vorzugreifen - nicht nur innerhalb des Projektes CEEC, bei dem zwei Mitarbeiter auch zu zwei verschiedenen Interpretationen einzelner Definitionsteile kamen, sondern auch bei einem Harmonisierungsversuch mit einem weiteren

[9] Dadurch können weitere 'Meta-Informationen' an die eigentlichen Inhalte angelagert werden: Auflösungen und Normalisierungen; Typbeschreibungen z.B. in welcher Sprache die Informationen gegeben sind oder welcher Klasse ein bestimmtes Element in einer bestimmten Perspektive zugeordnet werden kann etc.

Projekt zur Retrokonversion von Handschriftenkatalogen an der bayerischen Staatsbibliothek: dort war man zu wiederum abweichenden Auffassungen gelangt.

Zur Veranschaulichung der Problematik mögen zwei konkrete Beispiele genügen: Erstens: Der "Inhalt" einer Handschrift wird innerhalb eines Elements "msContents" beschrieben. Die Erläuterung zu diesem Element lautet:[10]

> *"The <msContents> element is used to describe the intellectual content of a manuscript or manuscript part. It comprises either a series of informal prose paragraphs or a series of more structured <msItem> elements, each of which provides a more detailed description of a single item contained within the manuscript."*

Hier werden nicht nur zwei grundverschiedene Konzepte - beschreibender Fließtext und weiter strukturierte formale Unterelemente - zugelassen. Fallstricke lauern auch im Begriff des 'intellectual content'. Ob globale Bemerkungen zur Handschrift, die Beschreibung von Korrekturen, Glossen, Notizen, Miniaturen, Diagrammen nun zum 'Inhalt' einer Handschrift oder eher zu ihrer äußeren Form gehören, darüber läßt sich lange streiten und hängt nicht zuletzt auch von der Perspektive von Handschriftenbearbeiter, Katalog und Benutzer ab.

Ein weiteres, allein schon zwischen Köln und München strittiges Beispiel liefert der 'Beschreibungskopf' zu einer Handschrift, abgebildet durch das Element "msHeading":[11]

> *The <msHeading> element may simply hold a short summary title, `heading', or `tombstone' specifying a supplied title or heading applicable to the whole of a manuscript, or this may be complemented with other elements.*

Zwischen einem reinen Titel und einer ganzen Serie von detaillierten Elementen, welche die gesamte Handschrift betreffen, kann hier eine große Bandbreite unterschiedlicher Konzepte umgesetzt werden. Je nach Auffassung könnte eine insgesamt beschreibende Notiz zu einer Handschrift bereits jetzt einem der beiden genannten Elemente <msContents> oder <msHeading> zugeordnet werden.

Beispiele dieser Art ließen sich beliebig vermehren - entscheidend für die Tragweite der praktischen Divergenzen ist aber ein Blick auf die schließlichen Nutzungsformen der Daten: Bei der Ausgabe von Katalogisaten wird es passieren können, daß man eine allgemeine Notiz zum Handschrifteninhalt nicht unter der Inhaltsbeschreibung finden wird, weil sie in der Kopfzeile steht (oder umgekehrt) -

[10] Abschnitt 2.5 der formalen Referenz-Dokumentation: <http://www.hcu.ox.ac.uk/TEI/Master/Reference/ms.html#msco>.

[11] Abschnitt 2.4 der formalen Referenz-Dokumentation: <http://www.hcu.ox.ac.uk/TEI/Master/Reference/ms.html#msdo>.

dies bildet allerdings nur die Unterschiedlichkeit der traditionellen gedruckten Kataloge (und ihrer Konzepte) ab und fällt damit in den Bereich adäquater Benutzung. Auf der Seite der zweiten digitalen Zugriffsform - zum Lesen/Stöbern tritt das gezielte Suchen - bleibt der Gewinn gegenüber der Typographie aber erhalten: Ob ein Buchschmuck-Detail (in MASTER: <decoNote>) den Vorlagen oder Traditionen gemäß bei der äußeren Beschreibung, bei einer gesonderten kunsthistorischen Betrachtung oder bei der genaueren Aufschlüsselung des 'intellektuellen Inhalts' erwähnt und in den Daten ausgezeichnet wurde, spielt keine Rolle - gefunden wird es in jedem Fall.[12]

[MASTER-Standard:] die inhaltliche Perspektive

Ein Blick auf die obersten Gliederungsebenen des Standards mag die konkrete inhaltliche Struktur des Modells grob andeuten:

Abb. 1: Inhaltliche Grobstruktur MASTER-Standard

[12] Dies setzt voraus, daß gleichartige Konzepte auch dann in gleicher Weise formalisiert sind, wenn sie in unterschiedlichen übergeordneten Kontexten auftauchen. Eine Buchschmuck-Detail muß (in MASTER) immer <decoNote> heißen können (und braucht deshalb auch nur einmal definiert zu werden), egal ob es in der äußeren Beschreibung oder beim Handschrifteninhalt vorkommt.

Diese Datenstruktur weist über die dargestellten Elemente hinaus zahlreiche Unterelemente, Verästelungen und Verfeinerungen auf. Die weitere Ausdifferenzierung läßt sich unter drei Aspekten beschreiben:

1.)Es gibt eine weitere hierarchische Untergliederung der Elemente. Innerhalb der Beschreibung des Äußeren einer Handschrift (<physDesc>) gibt es das Element Buchschmuck (<decoNote>), das wiederum in verschiedene Buchschmuck-Details (<decoNote>) zerfallen kann. Dieser Prozeß ist rekursiv, d.h. einzelne Buchschmuck-Details (oder -aspekte) können selbst wieder eine Reihe von Buchschmuck-Details enthalten. Dies kommt in der Praxis durchaus vor, wenn in einem Katalog z.B. zunächst eine Reihe ganzseitiger Miniaturen und anschließend alle Zierinitialen beschrieben werden. Leicht nachvollziehbar dürften solche Strukturen auch bei der Beschreibung des Inhaltes sein: Eine Handschrift enthält mehrere Texte, die wiederum aus mehreren Abschnitten bestehen können.

   Der Kennzeichnung der Elemente (den Auszeichnungselementen) können Zusatzangaben (Attribute) beigegeben werden. Auf diese Weise können z.B. Typologien gebildet werden oder die Elemente weiter spezifiziert werden, um sie für die spätere Verwendung zu differenzieren. Ein Beispiel hierfür wäre das Element <dimensions>, mit dem Größenangaben ausgezeichnet werden. Das Element kann offensichtlich in verschiedenen Zusammenhängen (bei einer Initiale ebenso wie beim Beschreibstoff oder dem Einband) sinnvoll sein, bedarf dann aber ggf. auch einer Spezifikation, auf was sich die Größenangabe bezieht.[13]

2.)Die Elemente enthalten 'Text'. Dieser kann nicht nur in einer hierarchischen Struktur weiter ausgezeichnet werden. Es gibt daneben eine Vielzahl von Elementen, die in verschiedener Funktion fast überall in solchen "Texten", d.h. unter Umständen auch innerhalb von beliebigen Elementen, auftauchen können. Hier sind z.B. Layout-Anweisungen zu nennen, mit denen Absätze getrennt, sinnvolle Zeilenumbrüche erzwungen oder Hervorhebungen markiert werden. Andere Auszeichnungen werden verwendet, um später automatische Verweise (Hyperlinks) zu erzeugen; das naheliegendste Beispiel dafür ist die Kennzeichnung von Seitenangaben der beschriebenen Handschrift durch das Element <locus>, durch das eine automatische Verlinkung mit z.B. Abbildungen der jeweiligen Seite möglich gemacht werden. Elemente der Art <person>, <place> oder <date> können schließlich dazu genutzt werden, bestimmte Begriffe so zu kenn-

[13] In der Praxis würde die Auszeichnung dann z.B. <dimensions type="leaves"> oder <dimensions type="initial"> lauten.

zeichnen, daß sie automatisch in ein Register eingehen oder für kombinierte Suchanfragen verwendet werden können.

CEEC und MASTER

[CEEC und MASTER:] Grundstrategie und Bedingungen

Bei der Beschreibung des MASTER-Standards ist bereits an einigen Stellen die Praxis des Projektes CEEC eingeflossen. Im folgenden wird die Verwendung des Standards und die tatsächliche Vorgehensweise bei der Realisierung einer 'virtuellen Handschriftenbibliothek' auf der Basis diese Standards kurz umrissen.

Innerhalb von CEEC wird der MASTER-Standard als äußere Richtlinie verwendet. Das impliziert zum einen, daß die weiteren Elemente des MASTER-Projekts, wie Softwarehilfen und ähnliches für die digitale Kölner Handschriftenbibliothek keine Rolle spielen und bedeutet zum anderen, daß eine institutionelle Kooperation (vorerst) nicht stattfindet.

Der MASTER-Standard beschreibt Handschriftenkatalogisate, also gewissermaßen die 'Metadaten' zu den Handschriften. Die Codices Electronici Ecclesiae Coloniensis zielen in einer mittelfristigen Perspektive aber auch auf eine tiefere Erschließung und potentiell auf die Aufbereitung der eigentlichen Handschriftenineinhalte. Angestrebt ist deshalb eine Datenstruktur, die neben den Metadaten die Daten (die Inhalte) abbilden kann. In diesem Sinne bildet der MASTER-Standard den Kopf einer größeren, weiter strukturierten Datei, die potentiell alle Informationen zu einer Handschrift integrieren können soll. Diese Gesamtdatei kann dann neben den Metadaten (den Kataloginformationen) auch die Bilddigitalisate[14] oder die Texte der Handschrift[15] beinhalten. Da die Gesamtdatei ebenfalls auf der Grundlage von XML durch eine erweiterte DTD beschrieben wird und für die Erweiterungen nach Möglichkeit bereits vorhandene Elemente aus

[14] Die Datei enthält nicht die Bilder selbst, sondern eine vollständige Konkordanz zwischen der 'gebräuchlichen' Benennung der Seiten eines Kodex (also z.B. "Dom Hs. 75, fol. 134r") und den Dateinamen der Bilder (also z.B. "kn28-0075-0268.jpg"). Diese Konkordanz wird dann von der Datenbank auch verwendet, um Seitenangaben im Katalogisat direkt in Hyperlinks zu den Abbildungen zu verwandeln.

[15] Mit der Wiedergabe der Texte wird der erste Schritt in Richtung 'Edition' oder 'editionsähnlicher Formen' unternommen. Die damit verbundenen theoretischen Implikationen, die Praxis der Integration von Transkriptionen oder aus gedruckten Editionen gewonnenen Textrepräsentanzen und die noch zu entwickelnden Modelle für die formale Abbildung 'sämtlicher' optischer Befunde einer Textseite in Hinsicht auf Paläographie/Graphematik, Textstruktur und Layout/Anordnung bleiben vorerst späteren Projektabschnitten vorbehalten.

den Standards der TEI benutzt werden, entsteht zwischen den einzelnen Teilen kein technischer oder konzeptioneller Bruch.

'Verwendung' des MASTER-Standards bedeutet innerhalb von CEEC grundsätzlich eine möglichst nahe Anlehnung, nicht aber einfache Übernahme. Haupterfordernis im Projekt ist es, die Informationen der vorhandenen Kataloge zur Dom- und Diözesanbibliothek möglichst vollständig und ohne Reibungsverluste durch eine Transformation in elektronische Formate verfügbar zu machen. Dazu werden in der Datenstruktur vor allem die Konzepte abgebildet, die hinter der Katalogisierung standen. Soweit sie sich auch in MASTER finden, wird MASTER übernommen; wo es abweichende Konzepte in den Katalogen gibt, wird der MASTER-Standard für die Kölner Situation verändert.

Ein prinzipieller Unterschied zwischen CEEC und MASTER besteht darin, daß innerhalb des CEEC-Projekts weder ein *neuer* Katalog *gemacht* wird, noch überhaupt *ein* Katalog erstellt wird. Es werden zunächst keine neuen Erkenntnisse zu den Handschriften gewonnen und die vorhandenen werden auch nicht in dem Sinne integriert, daß parallele (und sich widersprechende) Informationen zu eindeutigen Aussagen zusammengeführt würden. Der neuere Katalog ersetzt nicht den älteren, die ausführlichere Darstellung nicht die Kurzangabe. Die Autorität aller Vorlagen bleibt unangetastet.

Damit werden Informationsverluste - und sei es nur aus einer wissenschaftshistorischen Perspektive - vermieden, zugleich werden eine Reihe konzeptioneller bzw. technischer Probleme offensichtlich, wenn zahlreiche unterschiedliche Kataloge in *ein* Modell integriert werden müssen, ohne selbst desintegriert werden zu dürfen. Ein Überblick über die verschiedenen Katalogtypen im Projekt mag das Dilemma aufzeigen. Es gibt - jeweils nur für Teilbestände der Kölner Handschriften! - :

- Moderne ausführliche handschriftenkundliche Kataloge
- Moderne Kurzkataloge (Zensus)
- Ältere, unterschiedlich ausführliche Kataloge, teils aus historischer, teils aus bibliothekarischer Sicht.
- Paläographische Literatur mit katalogartigen Beschreibungen
- Kunsthistorische Literatur mit katalogartigen Beschreibungen
- Philologische Literatur mit katalogartigen Beschreibungen
- Ausstellungskataloge mit kunsthistorisch orientierten Beschreibungen

Insgesamt existieren knapp 20 Kataloge und Publikationen mit katalogartigen Beschreibungen. Die Komplikationen bei der Anwendung eines einheitlichen Standards ergeben sich zusammenfassend aus:

- unterschiedlichen fachlichen Perspektiven
- verschiedenen Altersstufen (Aktualität und Gültigkeit der Informationen!)
- unterschiedlicher Ausführlichkeit und
- verschiedenen Textsprachen[16].

Bevor der Arbeitsprozeß zur Erstellung des Datenmodells und seiner schließlichen Inhalte skizziert wird, soll hier nur kurz auf die Lösung für das Problem paralleler Informationen eingegangen werden. Potentiell allen Elementen kann im CEEC-Projekt ein Attribut 'Autorität' beigegeben werden; dieses enthält Informationen über die Herkunft einer Information, im Regelfall z.B. Autor (als Kennwort einer Publikation) und Seitenzahl. Gleichzeitig existiert eine Rangfolge innerhalb der verschiedenen Autoritäten, so daß z.B. zur Erstellung eines verkürzten Katalogisats die konkurrierenden Informationen mit geringerer Autorität ausgeblendet werden können. Dieses Konzept macht widersprüchliche Angaben nicht nur verwaltbar, es öffnet den elektronischen Katalog auch für punktuelle Ergänzungen und Erweiterungen, wie sie z.B. von externen Spezialisten eingebracht werden können. Solche 'Beiträger' erhalten ebenfalls einen Platz in der Autoritätenfolge, so daß ihre Informationen ggf. ältere oder weniger glaubwürdige Aussagen in einer Kurzdarstellung überdecken könnten.[17] Es bleibt das Problem, daß bei einer Vergesellschaftung paralleler Informationen in den detaillierten Verästelungen des 'Elementbaumes' der Fließtext eines gedruckten Katalogisats 'zerschnitten' und damit in seiner sprachlichen linearen Konsistenz desintegriert würde. Innerhalb von CEEC besteht die Parallelität der unterschiedlichen Informationsquellen deshalb auf einer sehr frühen Stufe: Statt *einer* integrierten Beschreibung der äußeren Merkmale mit vielen parallelen Angaben zum Layout bestehen *mehrere* Beschreibungen der äußeren Merkmale. Innerhalb dieser Beschreibungen bleibt der Text der einzelnen Kataloge unzerschnitten bestehen und damit lesbar.[18]

[16] Für die Kölner Handschriften liegen zur Zeit katalogartige Beschreibungen in Deutsch, Latein, Englisch, Französisch und Holländisch vor.

[17] Selbstverständlich bleiben *alle* Informationen in einer vollständigen Darstellung immer sichtbar.

[18] Da eine gewisse Veränderung der ursprünglichen Textstruktur bei der Überführung in das Datenmodell nicht immer zu verhindern ist, werden teilweise noch gesonderte HTML-Fassungen der Katalogisate bereitgestellt, die ein *Nachlesen* des Gesamtzusammenhanges erlauben.

[CEEC und MASTER:] Von der globalen zur lokalen DTD

Um aus der von MASTER vorgeschlagenen DTD zu einer spezifischeren Definition zu kommen, die den speziellen Bedingungen der Kölner Kataloge und des Kölner Projektes gerecht wird, wurde folgende Vorgehensweise gewählt, die in ihrer Abfolge chronologisch ist:

1. Entwicklung eines allgemeinen Modells in Anlehnung an den MASTER-Standard (d.h. auch semantische Interpretation der Beschreibung), zugleich aber unter Berücksichtigung anderer Standards (eBind[19], EAD[20], TEI) und spezieller Kölner Projektziele (zur Erinnerung: Katalogdaten als Teil einer Gesamtrepräsentation einer Handschrift; integrierte Verwaltung vollständiger digitaler Faksimiles).
2. Überführen der Katalog-Informationen in XML-Dateien. Dabei waren unsere Interpretationen der Standard-Beschreibung ebenso ständig zu verändern, wie spezielle Konzepte in den Katalogen zu neuen formalen XML-Elementen umformuliert werden mußten.
3. Verengung der DTD auf den realen Stand der Daten, d.h. auf die realen Konzepte der Kataloge. Die DTD bildet zu diesem Zeitpunkt genau die Struktur ab, die in den Daten vorhanden ist. Es beginnt dann ein zweiter Verengungs-Prozeß bei dem semantische Prüfungen der DTD immer wieder mit der formalen ('grammatikalischen') Validierung der Daten mittels der DTD und der Kontrolle der Semantik der XML-Daten einhergehen.[21]
4. Ggf. Rückbindung an das MASTER-Projekt. Falls sich in Köln neu entwickelte Elemente, Attribute oder Typologien als sinnvoll erweisen sollten, wäre eine Übernahme in den allgemeinen MASTER-Standard zu prüfen.

   Eine möglichst enge DTD sorgt für eine hohe Konsistenz der Daten. Sie erlaubt die fortgesetzte Prüfung dieser Konsistenz bei der Einfügung weiterer Informationen, sei es durch weitere Kataloge, durch Projekt-interne Ergänzungen oder durch externe Beiträge. Gleichzeitig muß sie fortlaufend verändert werden, wenn das Strukturkonzept und die Auszeichnungssemantik im Laufe der Zeit immer weiter verfeinert werden. Die DTD selbst ist auf den Internetseiten des Projekts dokumentiert und so auch von außen einsehbar.

[19] Die offizielle, allerdings recht alte Startseite zum Projekt eBind: <http://sunsite.berkeley.edu/Ebind/>

[20] Die offizielle Internet-Seite zur Encoded Archival Description (EAD): <http://www.loc.gov/ead/>.

[21] Etwas konkreter: Eine DTD kann daraufhin 'gelesen' - und verändert! - werden, ob man die abgebildeten Strukturen für 'sinnvoll' hält. Eine anschließende Validierung der Daten mit der DTD zeigt die Stellen, die mit der modifizierten Strukturbeschreibung nicht mehr übereinstimmen. Es ist dann abzuwägen, ob - je nachdem, was vernünftiger erscheint - Veränderungen in den Daten vorzunehmen oder die Änderung der DTD zurückzunehmen ist.

Aus einer allgemeinen Perspektive ist über die Verwendung des MASTER-Standards innerhalb der CEEC damit alles gesagt, es folgen nun Anmerkungen zu einigen eher konkreten, praktischen Fragen.

[CEEC und MASTER:] Praxis der DTD-Entwicklung und -Verwendung

Bei der Verwendung und Anpassung des MASTER-Standards tauchen in der Praxis Probleme auf, die zwei Bereiche betreffen: die Erweiterungen und Veränderungen an der DTD und den einzuspeisenden Daten einerseits und die Frage nach der Integrität von textlichen Informationen bei ihrer Überführung in formale Datenstrukturen andererseits.

Die Arbeit an der DTD umfaßt vier Arten von Veränderungen, die zu zwei bedenkenswerten Effekten führen:

1. Neue Elemente.

   Hoch spezialisierte und ausführliche Kataloge bilden unter Umständen Konzepte ab, die detaillierter sind, als das von MASTER repräsentierte Metakonzept. Um Informationsverluste gegenüber dem gedruckten Katalog zu vermeiden ist deshalb die Einführung neuer Elemente unvermeidbar.[22]

   Die Idee, alle möglichen oder zumindest vorkommenden Konzepte der Katalogisierung wären bereits in MASTER enthalten oder könnten ohne weiteres dort eingefügt werden, stößt dort an Grenzen, wo Konzepte nicht vollständig formalisierbar sind: Kataloge enthalten in der Realität oft vermischte Anmerkungen zu allem, was dem Katalogisator *irgendwie bemerkenswert* erschienen ist.[23] Will man eine aussagelose Residualkategorie <varia variorum> vermeiden, müßte man solche Informationen in logische Teilkonzepte zerlegen, was nicht vollständig möglich sein kann, da ja bei der Erarbeitung der Katalogdaten normalerweise kein bis ins Letzte konsistentes und formalisiertes Modell hinter den Arbeiten gestanden hat.

2. Andere Regeln.

   Die DTD schreibt vor, welche Elemente wo verwendet werden dürfen. Dies ergibt ein logisches Modell, das der Realität der gedruckten Kataloge nicht immer

[22] Im Falle von CEEC enthält z.B. der Ausstellungskatalog "Glaube und Wissen im Mittelalter", München 1998, durchgängig Anmerkungen dazu, wie die Seiten einer Handschrift für die Beschriftung vorbereitet worden sind. Zur Übertragung dieser Informationen ist vorläufig das Element <preparationOfPage> eingeführt worden.

[23] Dazu gehören alle Arten von Auffälligkeiten oder Abweichungen vom Normalen. Aber auch Bezüge zu anderen Handschriften oder Teiltranskriptionen bemerkenswerter Stellen, Notizen oder Glossen sind häufig.

genau entsprechen kann.

Ein Beispiel: Der Buchschmuck war bei MASTER ursprünglich Teil der Beschreibung des Äußeren eines Kodex (<physDesc>). De facto werden aber Buchschmuck-Details, wie z.B. auffällige Zierinitialen, oft bei der Beschreibung der einzelnen Textteile und Textanfänge erwähnt. Innerhalb von CEEC war deshalb eine Aufweichung der Regel, an welcher Stelle <decoNote> vorkommen darf, nötig. Der MASTER-Standard selbst ist aber ebenfalls weder statisch noch abgeschlossen: ähnliche Veränderungen der Regeln sind auch dort in letzter Zeit - und z.B. für den genannten Fall - vorgenommen worden.

3. Andere Bezeichnungen für Elemente und Konzepte.

Die Schaffung formaler Regeln ist eine mehr oder weniger willkürliche Adaption oft unscharfer Konzepte, mit denen wir unsere reale Welt strukturieren.

Überall in Texten und Beschreibungen können z.B. 'Namen' und 'Personen' vorkommen, ein offensichtlich vages Konzept, das auf vielerlei Arten umgesetzt werden kann. Wegen seiner Universalität lag die Anlehnung an bereits bestehende Standards (hier: TEI) nahe. Im Ergebnis können vorkommende Personen nach MASTER mit den Elementen <name type="person">, <person> und <persName> beschrieben werden, wobei dahinter jeweils unterschiedliche inhaltliche Konzepte mit interpretationsfähiger sprachlicher Bestimmung (s.o.) stehen.[24] Innerhalb von CEEC wurde statt dessen zunächst ein globales <person>-Element verwendet, das im weiteren Verlauf immer noch untergliedert und ggf. den Vorstellungen bei MASTER angenähert werden kann.

4. Weitere Verfeinerungen der DTD (Attribute, Typologien, normalisierte Daten)

Der MASTER-Standard ist 'nach unten offen'. Die Schaffung weiterer spezialisierter Attribute und Typologisierungen innerhalb bestehender Elemente wird ausdrücklich befürwortet.

Als Beispiel seien hier zunächst Überlegungen angeführt, einzelne Abschnitte der inhaltlichen Beschreibung einer Handschrift (<msItem>) weiter zu spezifizieren. Es stellt sich dann aber auch die Frage, ob nicht z.B. die vielfältigen Buchschmuck-Details durch eine umfassende Typologie tiefer erschlossen werden könnten. Hier böte sich der Rückgriff auf bereits bestehende Modelle und Vorschläge an, wie sie z.B. - in rudimentärer Form - die "Richtlinien Hand-

[24] Zu <person> und <perName> siehe Absatz 2.2.4 der formalen MASTER-Referenz-Dokumentation (<http://www.hcu.ox.ac.uk/TEI/Master/Reference/ms.html>), zu <name> siehe den entsprechenden Abschnitt in Anhang A (Refererenz-Dokumentation für Elemente und Klassen) der gleichen Dokumentation (<http://www.hcu.ox.ac.uk/TEI/Master/Reference/ref/NAME.html>).

schriftenkatalogisierung" der DFG darstellen.[25]

Zusätzliche Attribute sind auch dann nützlich, wenn normalisierte Daten den Zugriff auf die vorhandenen Informationen verbessern sollen. Innerhalb der CEEC ist z.B. die Einfügung von bibliothekarischen Normansetzungen für Personen, Orte und Institutionen geplant. Dies wären dann formal gesehen Attribute zu den tatsächlich vorkommenden Bezeichnungen. Suchfunktionen und Registererstellung würden dadurch erheblich verbessert.

Effekt 1: Divergierende DTDs bei MASTER und CEEC

Wenn in Projekten zur Handschriftenerschließung abweichende DTDs verwendet werden, stellt sich die Frage nach der Kompatibilität und Interoperationalität innerhalb eines - zukünftigen - Metasystems. Hier sind verschiedene (hypothetische) Modelle denkbar: Ein Metasystem auf 'kleinstem gemeinsamen Nenner' würde spezialistische Sonderentwicklungen einzelner Projekte ignorieren und damit gewisse Informationsverluste zugunsten einer einfachen Realisierbarkeit des Gesamtsystems in Kauf nehmen. Eine Berücksichtigung aller lokalen Abweichungen und neuer Verfeinerungen wäre aufwendig, würde aber den Zugriff auf alle aufbereiteten Informationen gewährleisten.

Dies sind Spekulationen zu zwei Eckpunkten. Für die Praxis dürfte bedeutsamer sein, daß das Grundmodell der Daten einigermaßen einheitlich gehandhabt wird und somit ein gemeinsamer Zugriff auf die wichtigsten Informationen auch über verschiedene Projekte hinweg möglich wird. Was darunter an Spezialisierung geschaffen wird, kann nötigenfalls auf einer Metaebene ignoriert und im Rückgriff auf eine lokale Ebene immer noch genutzt werden.

Effekt 2: Offenes Ende

Weder innerhalb des MASTER-Standards, noch bei konkreten Erschließungsprojekten muß die Arbeit an der DTD jemals abgeschlossen sein. Bei stabil bleibender Grundstruktur ist eine weitere Verfeinerung im Detail nicht nur möglich, sondern auch wünschenswert, verbessert sie doch die Erschließungstiefe wie die Benutzbarkeit der Informationen. Fraglich ist nicht ob, sondern wo die Erschließung der Informationen weiter vorangetrieben werden sollte. Innerhalb des CEEC-Projektes wird dabei die Prioritätensetzung von den Wünschen oder der Mitarbeit der Benutzer abgeleitet.

[25] Z.B. Richtlinien Handschriftenkatalogisierung, herausgegeben von der Deutschen Forschungsgemeinschaft, Unterausschuß für Handschriftenkatalogisierung, Bonn-Bad Godesberg [5]1992, S. 31f.

Bei der Umsetzung gedruckter Texte in eine formale Datenstruktur bleibt das Problem, daß ein gewisser 'Rest-Widerspruch' zwischen der Integration der Informationen in die Datenstruktur und der damit möglicherweise verbundenen Desintegration des fließenden Textes gibt. Selbst bei einem so 'textnahen' Modell wie dem MASTER-Standard stellen diese Restfälle[26] unangenehme Gefahrenquellen für Informationsverluste durch Entkontextualisierung bzw. Fragmentierung der einzelnen Angaben dar. Innerhalb der CEEC wird deshalb angestrebt, die Kataloginformationen nicht nur in die Datenstruktur einfließen zu lassen und sie damit als lesbare Texte teilweise aufzulösen, sondern sie zusätzlich parallel als geschlossene digitalisierte Texte anzubieten, so daß der Rückgriff aus den fragmentierten Daten in den ursprünglichen Text möglich bleibt. Dies erfordert allerdings - trotz der im folgenden zu beschreibenden mehrstufigen Praxis der Katalogverarbeitung - einen zusätzlichen Arbeitsaufwand, der nicht in allen Fällen geleistet werden wird.

[CEEC und MASTER:] Praxis der Datenaufbereitung

Wie kommen die Informationen in den Computer? Wer überführt die gedruckten Vorlagen in die beschriebene Datenstruktur? Wie werden die Auszeichnungen vorgenommen? Grundlage aller Arbeiten sind - wie bereits erwähnt - knapp 20 Publikationen, die auf fast 1000 Seiten Angaben zu den Kölner Handschriften enthalten. Zum gegenwärtigen Zeitpunkt (Projektmonat 10) sind davon ca. 60% 'verarbeitet', d.h. in rudimentärer Form ausgezeichnet und in die Datenbank eingefügt. Der Datenbestand umfaßt dabei Kataloginformationen mit ca. 300.000 Wörtern. Diese sind durch 62.000 (Auszeichnungs-Elemente) mit zusätzlichen 66.000 Attributen erschlossen.

Da 'Auszeichnung' die explizite Kennzeichnung impliziter semantischer Strukturen und Befunde bedeutet, ist 'Handarbeit' nicht immer zu vermeiden. Sie wird auch bei der weiteren Verfeinerung eine immer größere Rolle spielen müssen. Für den Anfang stellte sich aber die Frage, wie der Auszeichnungsprozeß nach Möglichkeit automatisiert werden konnte. Ausgangspunkt dafür ist die Feststellung, daß die Grundstruktur eines gedruckten Textes sich in seinem Layout widerspiegelt und durch bestimmte syntaktische Muster (z.B. Schlüsselbegriffe) gegliedert wird.

[26] Ein konkretes Beispiel hierfür sind Kataloge, bei denen die Beschreibung der allgemeinen Schrift des Kodex' nahtlos übergeht in die der besonderen oder Auszeichnungsschriften, der Initialen und Diagramme, bis hin zu den Miniaturen oder ganzseitigen Buchmalereien. Die nach dem MASTER-Standard erforderliche Trennung von Schrift (<msWriting>) und Buchschmuck (<decoration>) zwingt hier zu problematischen Grenzsetzungen und u.U. der Aufspaltung eines grammatikalisch geschlossenen Satzes in zwei - in der Datenstruktur - räumlich getrennte Abschnitte.

Dies nutzend, sind ca. 90% der jetzt vorhandenen Auszeichnungen maschinell erstellt worden.[27]

Grundlage jeder automatischen Auszeichnung ist ein in elektronischer Form vorliegender Text.[28] Je reicher an Formatierungsmerkmalen dieser ist, um so leichter und um so tiefer kann er maschinell ausgezeichnet werden. Am elektronischen Text ansetzend wird im MASTER-Projekt die Verwendung einer spezialisierten Editor-Software vorgeschlagen, mit dem der menschliche Bearbeiter die Auszeichnungen im Text vornehmen sollte.[29] Diese Strategie kommt im CEEC-Projekt nicht zum Einsatz. Statt dessen werden - soweit dies möglich ist - Formatierungs- und Textmuster in Auszeichnungselemente übersetzt.

Zum Verständnis sei der Prozeß am Beispiel eines realen Kurzkatalogisats aus dem 'Handschriftencensus Rheinland'[30] kurz angedeutet:

Abb. 2: Beispielkatalogisat aus dem Handschriftencensus Rheinland

966 Sign.: Cod. 3.

Origines: Opera selecta
Pergament, 182 Bll., 24,2 x 18 cm
9. Jh.

- fol. 1r-94r Homiliae XVI in Genesim; fol. 94v leer; fol. 95r-182v Homiliae XIII in Exodum -

Cum dicit de salvatore et rogabant eum daemonia ... - ... Explicit Liber Exodi

Textausgabe: GCS 29. Origines Werke. Bd. 6, S. 1-279.

Lit.: Jaffé/Wattenbach, S. 2.

576

Der Transformationsprozeß verläuft möglichst von den eindeutigen und sicheren Merkmalen zu den uneindeutigen und von den Elementen auf höherer Ebene

[27] Zu dem Prinzip der automatischen Umsetzung von Layoutstrukturen in explizite semantische Kennzeichnungen vgl. auch unseren Beitrag (Patrick Sahle, Torsten Schaßan:) "Das Hansische Urkundenbuch in der digitalen Welt", Hansische Geschichtsblätter 118 (2000), S. 133-155.

[28] Die Notwendigkeit eines elektronischen Textes erleichtert auch die oben angesprochene Bereitstellung eines parallelen vollständigen Textes z.B. in HTML oder PDF.

[29] Siehe <http://www.cta.dmu.ac.uk/projects/master/notetabhelp.html>.

[30] Handschriftencensus Rheinland: Erfassung mittelalterlicher Handschriften im rheinischen Landesteil von Nordrhein-Westfalen. Hrsg. von Günter Gattermann. Bearb. von Heinz Finger, Marianne Riethmüller u.a. Wiesbaden 1993.

zu den feineren Verästelungen. Im vorliegenden Fall bietet das Muster "Sign.: Cod." einen eindeutigen Anhaltspunkt dafür, daß hier die Signatur genannt wird, so daß der Computer die Information ohne weiteres in ein entsprechendes (Schlüssel-)Element umwandeln kann. Davon ausgehend ist eine Zahl vor diesem Element (der Signatur) die Katalognummer innerhalb der hier auszuwertenden Publikation 'Handschriftencensus' (<idNo type="catNo">[31]) und die folgende Zeile mit fett gedrucktem Text der Titel der Handschrift. Ist dieser erst als solcher erkannt, kann durch das Auftreten oder Fehlen eines Doppelpunktes noch zwischen einem allgemeinen Titel und einer Autor-Werktitel-Kombination (die durch den Doppelpunkt weiter gegliedert wird) unterschieden werden. Auf diese Weise können fast alle Informationen auf höherer Ebene erkannt werden: Ein kursives *"Lit.:"* indiziert eine Literaturliste, das kursive Wort *"Textausgabe"* kennzeichnet Editionen, der kursive Text vor diesen Elementen enthält Incipit und Explizit, die genaue Grenze zwischen den beiden Elementen wird durch das Muster *"... - ..."* eindeutig bestimmt. Nach der Datumsangabe (ebenfalls eindeutig und sicher, weil für jede Handschrift angegeben) folgt die Inhaltsangabe, die durch Semikolon in einzelne Teile gegliedert wird. Die Seitenangabe in unserem Beispiel schließlich, die eindeutig immer als letzte dreistellige Zahl vor einem Seitenumbruch aufzufinden ist, wird herausgezogen, um zu den Elementen auf höchster Ebene die Herkunft der Informationen zu speichern.

Das System funktioniert so zuverlässig, wie die Struktur der Beschreibungen gleichmäßig ist. Von der Norm abweichende Beschreibungsteile kommen immer wieder vor, manuelle Kontrollen sind deshalb unvermeidlich, lassen sich aber ebenfalls maschinell vereinfachen: Eine Strategie ist die listenförmige Ausgabe aller gleichartigen Elemente. Das menschliche Auge ist sehr gut in der Lage in einer langen Liste von Textstücken, die strukturell gleichförmig sein sollten, 'Ausreißer' zu erkennen. Selbst Hunderte von Fällen sind in wenigen Sekunden zu überblicken.[32]

Die weitere Verfeinerung der Auszeichnungen kann dann später auch auf die Verarbeitung einzelner, bereits identifizierter Elemente beschränkt werden. Dies wäre z.B. der Fall, wenn solche Elemente mit anderen Wissensbasen oder Listen von Schlüsselwörtern abgeglichen würden: Schriftbeschreibungen mit einem fachterminologischen Glossar, allgemeine Texte mit Verzeichnissen von Orts- und Per-

[31] Vollständig: <idNo type="catNo" authority="Zensus:576">966</idNo>, natürlichsprachlich lautet die Aussage dann: "Bei dieser Handschrift ist 966 eine Identifikationsnummer vom Typ Katalognummer; diese Angabe stammt aus dem Handschriftenzensus Rheinland, Seite 576".

[32] Es soll aber auch nicht verheimlicht werden, daß die einfache Veränderbarkeit der Daten eine gewisse Toleranz gegenüber 'letzten' Fehlern mit sich bringt: Jeder Benutzer ist schließlich ein potentieller Fehlermelder und trägt damit zur Verbesserung der Inhalte bei.

sonennamen, Buchschmuckdetails mit Typologien, die durch eindeutige Schlagworte ("Diagramm", "Initial", "ganzseitig + Miniatur") indiziert werden können.

Dies ist ein Faß ohne Boden. In welche Richtung weitere Erschließungsarbeiten an den bereits vorhandenen Daten gehen sollen, kann nicht zuletzt auch von den Wünschen der Benutzer abhängig gemacht werden. Die Nutzung externer Wissensbasen ist zudem eine Strategie, die noch nicht allzu weit entwickelt ist und deren Effizienz und Zuverlässigkeit in den weiteren Projektphasen noch zu evaluieren sein wird.