Patrick Sahle

Datenstandards in der Erschließung historischer Dokumente

aus:

Forschung in der digitalen Welt

Sicherung, Erschließung und Aufbereitung von Wissensbeständen

Herausgegeben von Rainer Hering, Jürgen Sarnowsky, Christoph Schäfer und Udo Schäfer

S. 29–42

Impressum

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.
Die Online-Version dieser Publikation ist auf der Verlagswebsite frei verfügbar *(open access)*. Die Deutsche Nationalbibliothek hat die Netzpublikation archiviert. Diese ist dauerhaft auf dem Archivserver der Deutschen Nationalbibliothek verfügbar.

Open access über die folgenden Webseiten:
Hamburg University Press – http://hup.sub.uni-hamburg.de
Archivserver der Deutschen Nationalbibliothek – http://deposit.d-nb.de

ISBN-10 3-937816-27-5 (Printausgabe)
ISBN-13 978-3-937816-27-2 (Printausgabe)
ISSN 0436-6638 (Printausgabe)

Produktion: Elbe-Werkstätten GmbH, Hamburg, Deutschland
http://www.ew-gmbh.de

Bildnachweis: Der Abdruck aller Abbildungen erfolgt mit freundlicher Genehmigung der Autoren bzw. des Autors des jeweiligen Beitrags.

Inhaltsübersicht

Datenstandards in der Erschließung historischer Dokumente

Patrick Sahle

1. Vorbemerkung

Der diesem Text zugrunde liegende mündlich-visuelle Tagungsvortrag hatte vor allem einführenden, orientierenden, fast schon Workshop-Charakter. Im Wechsel der situativen, kommunikativen und medialen Umstände wird er hier auf die wesentlichen theoretischen Grundlinien reduziert. Die einzelnen Standards werden nicht eingehend besprochen und auch das abschließende Praxisbeispiel wird nicht wiedergegeben: Anhand des Projektes *Zentrales Verzeichnis Digitalisierter Drucke* (zvdd) waren im mündlichen Vortrag die Auswahl von Standards, ihre Lokalisierung im vorgestellten allgemeineren Modell und die praktischen Schwierigkeiten ihrer Anwendung besprochen worden. Bei dem hier vorliegenden Text handelt es sich inhaltlich eher um ein Abstract, umfänglich um ein ausformuliertes, erweitertes Abstract.

2. Zur Bedeutung von Standards

Jeder Fall ist ein Sonderfall. Alle Gegenstände der historischen Überlieferung erfordern ihre je eigenen Weisen der Beschreibung und Erschließung. Jede Materialgattung, jede Überlieferungssituation und alle jeweils zu erwartenden Auswertungsinteressen könnten die anzuwendenden technischen Standards in einer Weise bestimmen, die für jedes Projekt zu einer individuellen Lösung führen müsste. Dieser strikt material- und situationsorientierte Ansatz muss aber derzeit aus einer ganzen Reihe von Gründen

immer mehr relativiert werden. Mindestens drei Prinzipien des gegenwärtigen Medienwandels haben als allgemeine Paradigmen weit reichende Auswirkungen auf alle Teilbereiche der konzeptionellen und praktischen Konfiguration der Erschließung. Die Rede ist hier

- vom inkrementellen Prinzip,
- vom Prinzip der Vernetzung und
- vom Prinzip der Transmedialisierung.

Was bedeuten diese Prinzipien für die ‚Erschließung', unter der in einem weiten Begriffsverständnis alle Arbeiten zusammengefasst werden können, die historisch überlieferte Materialien durch Wiedergabe, kritische Beschreibung und Informationsanreicherung für die weitere wissenschaftliche Verwendung vor- und aufbereiten?

Früher war ein Faksimile ein Faksimile, ein Katalog ein Katalog, ein Regestenband ein Regestenband, eine Edition eine Edition und eine wissenschaftliche Monographie eine wissenschaftliche Monographie. Abgeschlossene und für sich stehende Informationseinheiten. Abgeschlossene Publikationsformen, die zwar immer auch aufeinander verwiesen, sonst aber in keinem weiteren direkten Verhältnis zueinander standen. Heute – realistisch betrachtet aber in der allgemeinen Praxis wohl erst morgen – ist alles potentiell vernetzt. Das, was uns früher als expliziter Verweis erschien, wenn zum Beispiel eine Edition sich auf die in den Quellenstudien, Katalogen und Findbüchern nachgewiesenen Materialien bezog, hat heute den Status nur eines impliziten Verweises. Schließlich ist mit der realen Verlinkung zwischen digitalen Ressourcen bzw. mit der Einbindung von digitalen Datenfragmenten eine neue Form der expliziten Verweisung und Bezugnahme zur leitenden Form des Möglichen geworden.

Die einzelnen Formen der Erschließung bauen in einer veränderten Weise explizit aufeinander auf. Dies lässt die Grenzen zwischen den Publikationseinheiten, die medial bedingt waren aber konzeptionell gefasst wurden, verschwimmen. Die traditionellen Formen greifen über sich hinaus: Der elektronische Katalog schließt auf der einen Seite digitale Faksimiles ein und bildet auf der anderen Seite potentiell auch eine Plattform für die Forschungsinstrumente zur Tiefenerschließung der Überlieferung. Die wissenschaftliche Auswertung, die geschriebene Geschichte, inkludiert ihre Grundlagen und macht sich so potentiell in einer neuen Weise transparent.

Auf der einen Seite verschwimmen die Grenzen. Auf der anderen Seite müssen die einzelnen Arbeiten an der Überlieferung und die dabei entstehenden Informationen doch wieder konzeptionell getrennt werden. Sie müssen in einer neuen Weise ‚modularisiert' werden. Sie müssen als Bausteine in einem umfassenden Informationsraum gedacht werden, der jetzt als in sich geschlossener und zusammenhängender Informationsraum eben dadurch besser sichtbar wird, dass er nicht mehr fragmentiert, sondern potentiell hoch integriert ist.

Die ‚Trennungen' in der digitalen Erschließung sind teilweise andere Trennungen als in der traditionellen Erschließung. Zunächst ist für die digitalen Daten oft die Rede von der Trennung von Inhalt und Form. Dabei handelt es sich um ein weit verbreitetes Grundkonzept der gegenwärtigen technischen Lösungsansätze. Gemeint ist, dass Informationen auf einer Ebene abstrakter Datenmodelle und Daten gespeichert werden sollen. Dabei sollen aber die Inhalte von Informationen gespeichert und verwaltet werden und nicht ihre äußere Erscheinungsform in jenen Präsentationsmedien, mit denen der menschliche Benutzer schließlich in Kontakt kommt. Durch die Speicherung verallgemeinerter, codierter Informationen wird es außerdem möglich, dass gleiche Daten in verschiedenen medialen Konfigurationen mit unterschiedlichen Ergebnissen algorithmisch ausgegeben werden. Das Konzept beinhaltet so die Trennung von transmedialen Daten und (beliebigen) medialen Ausgabeformen. Da die medialen Ausgabeformen erstens algorithmisch erzeugt werden und zweitens höchst variabel sind, geht es vordringlich nur um die Daten, die ‚Inhalte' selbst. Dies ist die oben angesprochene ‚Transmedialisierung'.

Die transmedialisierten Inhalte können leichter explizit aufeinander aufbauen – sie können sich gegenseitig übernehmen und integrieren. Alle Arbeiten an der Überlieferung erscheinen heute als Bausteine einer Verarbeitungskette, die bei der unmittelbaren materiellen Überlieferung ansetzt und diese immer tiefer erschließt, kritisch sichtend, kontextualisierend, transkribierend, deutend, analysierend immer weiter verarbeitet. Damit führen aber auch alle Arbeiten zu Bausteinen der Erschließung – zu Bausteinen, deren Wert darin liegt, dass sie möglichst einfach in anderen Zusammenhängen benutzbar sein sollten.

Wenn man nun betrachtet, wie die einzelnen Bereiche der Erschließung differenziert werden können, dann sehen wir, dass zunächst durchaus von den traditionellen Konzepten ausgegangen werden kann, um zu einer neu-

en Modularisierung zu kommen. Offensichtlich gibt es verschiedene Formen der Grunderschließung, die von der Objektbeschreibung über die Katalogisierung bis zur Regestierung (gewissermaßen einer ersten inhaltlichen Durchdringung) reichen kann. Neu ist hier, dass noch ganz am Anfang die digitale Faksimilierung durch die Umkehrung der traditionellen Kostenrelationen[1] zu einer Selbstverständlichkeit wird, die zu einer anderen Transparenz der weiteren Erschließungsschritte führt und die relativen Positionen der einzelnen Module neu bestimmt.[2]

Auf die Grunderschließung setzt die Textgewinnung auf. Hier wird der Text zunehmend als Skala verschiedenster, im Idealfall algorithmisch und dynamisch zu generierender Textformen erkennbar, die zwischen den Polen quellennaher (benutzerferner) und quellenferner (benutzernaher) Darstellungsweisen eine Vielzahl von Positionen einnehmen können. Auswertende Formen der Arbeit mit der Überlieferung können in gleicher Weise auf einer Skala fortschreitender Verarbeitung und zunehmender Distanz zu den Quellen verortet werden. Auf dieser Verarbeitungskette können wir durchaus den gleichen Akteuren einen Platz zuweisen, die wir auch in der traditionellen Form spezielle Arbeiten durchführen gesehen haben: Archivare und Bibliothekare kümmern sich vor allem um Grunderschließungsleistungen. Wissenschaftliche Editoren erstellen verschiedene Formen von Textrepräsentationen. Und Fachwissenschaftler der unterschiedlichsten Richtungen werten das überlieferte Material schließlich für spezielle Fragestellungen aus. Allein die Übergänge werden jetzt vielleicht noch fließender als sie es bislang schon waren, da die integrative Kraft der transmedialen Daten die Grenzen zwischen den Publikationsformen dieser Arbeitsschritte verwischen.

Die Modularisierung der Arbeitsschritte geschieht auf der Achse der zunehmenden Verarbeitung der Überlieferung. Sie bezieht sich auch auf die Rolle der verschiedenen Akteure im Prozess der umfassenden Erschließung. Sie ergibt sich aber auch aus einem anderen Aspekt der gegenwärtigen Informationstechnologien: Vernetzung, Integration und Transmediali-

[1] Darauf war bereits Manfred Thaller in seinem Eröffnungsreferat eingegangen.

[2] In dem Moment, in dem das digitale Faksimile selbstverständlich als erstes verfügbar gemacht wird, verändert sich der ursprüngliche ‚Stellvertretercharakter' von Katalogisaten, Quellenbeschreibungen, Regesten und Textausgaben. Sie werden nun stärker als Informationsdestillate, als Hilfsmittel zum Auffinden von Quellen und als vorverarbeitende Formen für spezielle Auswertungsstrategien erkennbar.

sierung können nur dann funktionieren, wenn klar ist, welches die Objekte sind, die wir inhaltlich repräsentieren wollen. Die gegenwärtigen Informationstechnologien sind darauf angewiesen, mit klar definierten Informationseinheiten zu operieren. Auf unterschiedlich granularen Ebenen muss klar sein, wovon jeweils die Rede ist, was eigentlich der – zunächst konkrete, dann aber auch fortschreitend abstrakt gefasste – Gegenstand ist, von dem die Daten handeln. Auch deshalb ist eine klare Modularisierung unseres Informationsraumes unerlässlich. Wir müssen absehen können, über welche Objekte Informationen zu erwarten sind, wie wir diese (repräsentierten) Objekte ansprechen können und wie wir Informationen zu ihnen verstehen und verarbeiten können. Die Form definierter Objekte und der Zuschnitt der Informationen zu ihnen müssen berechenbar sein, um in anderen Kontexten benutzt werden zu können. Und sie müssen bekannt sein, damit Schnittstellen zwischen technischen Systemen funktionieren können, über die Daten weitergegeben werden können.

Die Definition von Objekten, Teilbereichen der Verarbeitung, Festlegung von Sichten auf Objekte und Schnittstellen ist auch deshalb so wichtig, weil die Organisation des Informationsraumes tatsächlich noch ein wenig komplexer ist, als dies bislang angedeutet worden ist. Es muss nämlich weiterhin unterschieden werden zwischen der Eingabeform von Informationen, der Form der Datenhaltung und der Ausgabeform. Erschließungsinformationen können auf verschiedenen Wegen gewonnen werden. Durch die Konversion bereits vorhandener Ressourcen ebenso wie über die Eingabe neuer Informationen. Die Schicht der Datenhaltung ist davon weitestgehend unabhängig. Zwischen beiden besteht ein Verhältnis der Übersetzung und Abstraktion. Auf der anderen Seite besteht ein Verhältnis der Übersetzung und medialen Konkretisation zwischen der Verwaltung allgemeiner Daten und ihren Darstellungsweisen in den verschiedenen medialen Konfigurationen. Erschließungsdaten können hier ebenso zu einem gedruckten Werk ausformatiert werden, wie mit ihnen unter Umständen eine interaktive Netzpublikation angetrieben werden kann. Was die Modularisierung des Informationsraumes betrifft, so können auf der Seite der Publikation verschiedene inhaltliche Module wieder zusammenfließen.

Eine weitere Differenzierung der Teilbereiche kann noch im Bereich der Datenhaltung vorgenommen werden. Hier ist – für manche technische Ansätze – noch zwischen Grunddaten, Strukturinformationen und Metadaten zu unterscheiden.

Der Informationsraum der Erschließung historischer Dokumente lässt sich insgesamt wie folgt skizzieren:

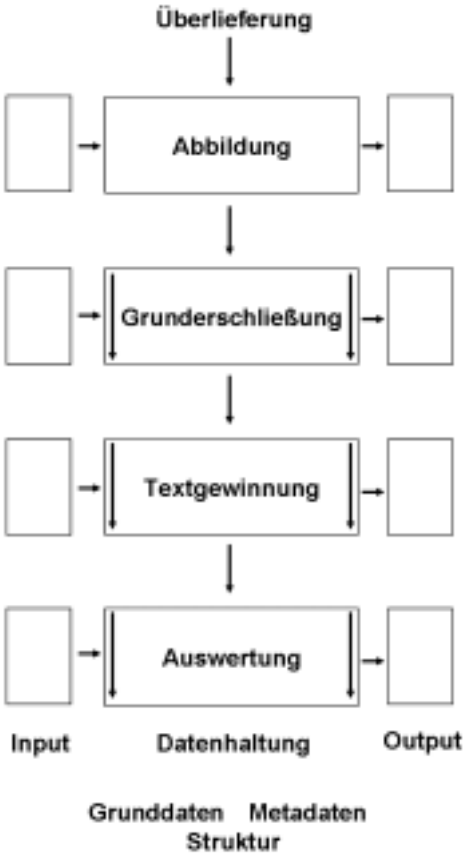

Abbildung: Erschließung der Überlieferung als Informationsraum.

Das Modell des Informationsraumes soll dabei helfen, die Bedeutung von konzeptioneller Modularisierung, klarer Objektdefinitionen, von Standards und von Schnittstellen in der Erschließung der historischen Überlieferung einsichtig zu machen. Es dient aber auch als Folie, um Standards zu lokalisieren bzw. allgemeiner, um einzelne Erschließungsprojekte und deren Verwendung von Standards zu beschreiben.

Die Bedeutung technischer Standards liegt in der Ermöglichung eines modularen Erschließungskonzeptes. Die Grundfunktion von Standards ist es, Informationen durch Gleichmäßigkeit der Beschreibung auch nach außen besser verstehbar und berechenbar zu machen. Diese Berechenbarkeit ist die Grundlage für die Erfüllung weiterer Anforderungen an digitale Erschließungsinformationen: für die Nachhaltigkeit und langfristige Nutzbarkeit der gewonnenen Daten, für die algorithmische weitere Verarbeitung in anderen Nutzungssituationen und für die Anschlussfähigkeit und Integrationsfähigkeit in weitere Kontexte. Denn die einzelnen ‚Bausteine' der Erschließung werden dadurch über ihre unmittelbare Erstellungssituation hinaus wertvoll, dass sie auch später noch in anderen Zusammenhängen nachgenutzt, weiter verarbeitet oder in andere Systeme integriert werden können. Dazu gehören dann zum Beispiel vertiefende Erschließungsarbeiten. Dazu gehören unter Umständen aber auch Suchverbünde oder Portale, die Teilinformationen herausziehen, um die erschlossenen Objekte von einer übergeordneten Warte aus nachzuweisen, besser auffindbar und besser nutzbar zu machen.

3. Standards im Informationsraum

Im Bereich der Standards begegnen einige scheinbare Paradoxien. So gibt es zwar einerseits bereits eine große Fülle von Standards, andererseits aber passt selten ein Standard ganz genau auf die konkret vorliegenden Probleme. Man soll die vorhandenen Standards benutzen und nicht immer wieder neue Standards erfinden, muss dann aber immer einen Mittelweg finden zwischen der glatten Anwendung bestehender Standards und den eigenen Bedürfnissen. Diese führen in der Regel zu Anpassungen, die zwar in so genannten ‚Anwendungsprofilen' dokumentiert werden können, letztlich aber den Status der verwendeten Datenbeschreibung ‚als Standard' gefährden. Dieses Problem resultiert daraus, dass Standards und die Situationen, in denen sie verwendet werden, durch eine ganze Reihe von Aspekten bestimmt werden, die noch über die Modularität des erschließenden Informationsraumes hinaus das Feld der Datenbeschreibungen ausdifferenzieren.

Für die Lokalisierung und Beschreibung von Standards, aber auch für die Auswahl zur Verwendung in bestimmten Projektkontexten, werden eine Zahl von Parametern genutzt, die im Folgenden aufgelistet werden.

3.1 Grad der Allgemeinheit

Standards können sich auf unterschiedlichen Abstraktionsebenen befinden. Das ‚relationale Datenmodell' oder das ‚hierarchische Datenmodell' sind allgemeine Weisen der strukturierten Abbildung von Informationen. Diese können sich in konkreteren Standards niederschlagen: SQL (Standard Query Language) ist ein Abfragestandard für Daten, die dem relationalen Modell folgen. XML (eXtensible Markup Language) ist eine Realisation des allgemeinen Modells ‚Textauszeichnung'. Dieses ist selbst sowohl an sequentiellen wie auch hierarchischen Grundstrukturen orientiert. XML ist ein Beispiel für einen Meta-Standard, der aus einer allgemeinen Warte zwar konkret ist (indem er ein bestimmtes Datenmodell realisiert und genaue Struktur- und Syntaxvorgaben macht), aus einer speziellen Warte aber allgemein. Für die konkrete Anwendung muss auf der Grundlage des Meta-Standards erst eine Anwendungssprache formuliert werden.

3.2 Determinismus

Standards sind wie Sprachen. Ihr Funktionieren liegt weniger in einer abstrakten Vorgabe, wie etwas gemeint ist, als vielmehr im gemeinsamen Gebrauch der Sprecher. In der Regel sind Standards nicht selbsterklärend, sondern müssen dokumentiert werden, um ihren Sinn zu klären. Weil diese Dokumentation aber natürlichsprachlich erfolgt, bleibt notwendigerweise immer Raum für abweichende Interpretationen bei den Benutzern. Dieser Raum kann unterschiedlich groß sein. Ein Standard kann zu sehr gleichmäßigen Anwendungen führen wie auch zu sehr ungleichmäßigen. Dies hängt nicht unbedingt von der Präzision und Vollständigkeit der Dokumentation ab, sondern auch von den anderen hier beschriebenen Parametern.

3.3 Beschreibung oder Austausch

Standards können die Beschreibung von Informationen festlegen oder einen technischen Rahmen für den Austausch von Daten bilden. Solche Schnittstellen-Standards können wiederum unterschiedlich allgemein sein.

Allgemeine Übertragungsprotokolle wie HTTP (Hypertext Transfer Protocol) oder Schnittstellen wie CGI (Common Gateway Interface) bilden nur einen allgemeinen Rahmen für die Kommunikation der Daten, ohne etwas über die Inhalte auszusagen. Speziellere Schnittstellen-Protokolle wie Z39.50 bzw. ZING (Z39.50 International Next Generation) oder OAI-PMH (Open Archives Initiative, Protocol for Metadata Harvesting) betreffen dagegen bestimmte Datenbereiche, hier zum Beispiel die Metadatenschicht.

3.4 Materialgattung

Offensichtlich brauchen verschiedene Materialien ihre eigenen Standards zur Beschreibung. Die TEI-Guidelines (Text Encoding Initiative) betreffen Volltexte, der MASTER-Standard die Codierung von Handschriftenkatalogen (Manuscript Access through Standards for Electronic Records).[3] EAD (Encoded Archival Description) bezieht sich auf Findmittel in Archiven. SVG (Scalable Vector Graphics) dient der Wiedergabe von ‚Bildern' (Skizzen, Diagramme), die sich über Vektoren beschreiben lassen.

3.5 Verarbeitungsgrad

In meinem Modell des Informationsraumes spielte die fortschreitende Verarbeitung eine entscheidende Rolle. Manche Standards betreffen eher Grunddaten, manche spätere Verarbeitungsstufen. MASTER ist ein Standard für die Katalogisierung von Text-Objekten, TEI ein Standard für die Wiedergabe der Texte *in* den Objekten.

3.6 Input – Datenhaltung – Output

Manche Standards überspannen den Dreischritt Eingabe – Verwaltung – Ausgabe. Oft ist aber die Verwendung unterschiedlicher Lösungen sinnvoll. Dabei besteht für die Verwendung von allgemeinen Standards auf der Ebene des Inputs noch die geringste Notwendigkeit: hier können durchaus lokale technische Lösungen verwendet werden, weil der Austausch und die weitere Verwendung der gewonnenen Daten ja erst auf der Ebene der Datenhaltung ansetzt. Allein in sehr hoch modularisierten Projekten mit zum Beispiel weit verteilter Datenerfassung ist die Verwendung von Standards

[3] MASTER ist allerdings in der letzten Fassung der TEI-Guidelines (P5) aufgegangen.

bereits bei der Dateneingabe wichtig. Ein Standard für die Erfassung von Daten ist XForms (XML-Formulare). Die meisten Standards betreffen die Ebene der Datenhaltung, da es hier um eine möglichst sachadäquate, differenzierte Beschreibung und Verwaltung von Informationen geht. Datenhaltungs-Standards sind oft zugleich noch relativ allgemein, so dass sie für eine unmittelbare Darstellung in Output-Formen gar nicht geeignet sind. Dies betrifft zum Beispiel XML im Allgemeinen oder konkrete XML-Auszeichnungssprachen wie die TEI. Auf der anderen Seite können solche allgemeinen Daten dann algorithmisch in Ausgabeformate überführt werden, die so präzise sind, dass sie von verschiedenen Medien unmittelbar dargestellt werden können. Hier wäre zum Beispiel auf TEX, PS (PostScript) oder PDF (Portable Document Format) zu verweisen, mit denen Druckmedien generiert werden können. Mit PDF, dann aber vor allem auch (X)HTML (Hypertext Markup Language), wären hier auch die wichtigsten Standards für den Output von Texten in digitalen Medien genannt. Ähnliche Unterscheidungen wären parallel zum Beispiel für den Bereich der Bilddaten zu machen: Hier ist TIFF ein Eingabeformat, LZW-komprimiertes TIFF ein Speicherformat und JPEG, GIF oder PNG Ausgabeformate.

3.7 Informationsdimensionen

Grunddaten, Strukturdaten, Metadaten, Semantik. Standards können sich auf unterschiedliche Ebenen der digitalen Repräsentation, auf unterschiedliche Sichtweisen auf Objekte beziehen. ASCII (American Standard Code for Information Interchange) und UNICODE betreffen die Zeichencodierung von Textdaten, Auszeichnungssprachen wie die TEI regeln die Beschreibung der Struktur von Volltexten. Die TEI umfasst dann aber auch schon den Bereich der Metadaten, der für andere Objekte von gesonderten Standards abgedeckt wird: Sollen zum Beispiel nur die Metadaten bibliographischer Einheiten erfasst werden, so stehen dafür bibliothekarische Formate wie MAB (Maschinelles Austauschformat für Bibliotheken) oder MARC (Machine Readable Catalogue) zur Verfügung. Ein wichtiger Standard, der zwar für digitale Ressourcen im Allgemeinen ‚zuständig' ist, sich hier aber ebenfalls ausdrücklich auf die Metadatenschicht beschränkt, ist DC (Dublin Core). Metadaten- und Strukturdaten digitaler Objekte deckt METS (Metadata Encoding and Transmission Standard) ab, auf die Strukturdaten gedruckter Daten beschränkt sich E-Bind (electronic Binding

DTD). Gerade in letzter Zeit hat sich jenseits der deskriptiven Metadaten noch eine Schicht der eher deutenden semantischen Daten etabliert. Auch bisher haben inhaltlich orientierte Beschreibungsstandards wie MIDAS (Marburger Informations-, Dokumentations- und Administrationssystem) für Objekte der Kunst oder ‚Iconclass' für die Klassifikation von Bildinhalten im Grunde eine semantische Deutung ihrer Gegenstände vorgenommen. Mit den Standards im Vor- und Umfeld des kommenden ‚Semantic Web' sind hier aber allgemeinere Formate entstanden, die semantische Inhalte und Bezüge von digital beschriebenen Objekten verwaltbar machen. Zu nennen wären hier zum Beispiel RDF (Resource Description Framework), XTM (XML Topic Maps) oder OWL (Web Ontology Language).

3.8 Konzeptionalisierung

Selbst wenn für einen bestimmten Anwendungsfall alle bereits genannten Parameter klar sein sollten, so führt dies nicht dazu, dass am Ende *ein* Standard als Mittel der Wahl zur Verfügung stehen würde. Oft würde man feststellen, dass für den speziellen Fall noch kein Standard entwickelt worden ist, so dass man auf eine Beschreibungsweise ausweichen müsste, die in dem einen oder anderen Parameter von den eigentlichen Anforderungen abweicht. Oft hat man aber auch die Auswahl zwischen mehreren Standards. Diese unterscheiden sich dann zum Beispiel in der Art und Weise, wie sie die Beschreibung der Objekte konzeptionalisiert haben. Ein Beispiel hierfür wären MAB und MARC: Beide betreffen die Metadaten zu bibliographischen Einheiten. Sie beschreiben sie aber durchaus mit leicht unterschiedlichen Konzepten.

3.9 Technische Umsetzung

Selbst wenn sonst alle anderen Parameter übereinstimmen sollten, so kann ein Unterschied zwischen verschiedenen Standards außerdem immer noch in der technischen Umsetzung auch gleicher Konzepte liegen. Dabei können die technischen Unterschiede Auswirkungen auf die Gewinnung, die Speicherung und Verwaltung, den Austausch und die Darstellung der Informationen haben und dadurch Kriterien für die Benutzung eher des einen oder des anderen Formates liefern. Ein Beispiel aus diesem Bereich wären die Standards zur Codierung von Bildern nach dem Bildpunkte-

Prinzip (‚bitmaps'). JPEG (Joint Photographic Experts Group), PNG (Portable Network Graphics), GIF (Graphics Interchange Format) oder TIFF (Tagged Image File Format) dienen alle der Codierung von Bildern nach dem gleichen Prinzip. Sie unterscheiden sich allein nach der Art der Datenkomprimierung, den Farbräumen und den Möglichkeiten zusätzlicher Dokumentation der Bilder. Bereits dies können aber Aspekte sein, die im jeweiligen Anwendungsfall den Einsatz eines bestimmten Formats nahe legen.

4. Probleme und Fehlerquellen bei der Verwendung von Standards

Die Auswahl und Anwendung geeigneter Standards ist nicht trivial. Standards können mehrere Felder abdecken, sich überschneiden, andere Standards beinhalten oder die Verwendung ergänzender Standards nahe legen. Weiterhin ist zu überlegen, ob und wie weit Standards für die lokalen Erfordernisse angepasst werden sollten. Bei der Auswahl und Anwendung von Standards treten oft ähnliche Probleme auf, werden häufig ähnliche Fehler gemacht. Auf einige sei hier abschließend hingewiesen.

4.1 Falsches Datenmodell

Zu den seltener werdenden Fehlern gehört, dass für Informationen, die nun einmal eine bestimmte Grundstruktur aufweisen, das falsche Grunddatenmodell gewählt wird. In der Vergangenheit ist es häufig vorgekommen, dass zum Beispiel hierarchisch organisierte Dokumentinformationen oder Textdaten mit einem relationalen Modell abgebildet wurden.

4.2 Nicht materialadäquater Standard

Manchmal werden Standards für allgemeiner gehalten als sie es tatsächlich sind. TEI ist ein Standard, der vornehmlich aus literaturwissenschaftlicher und linguistischer Sicht heraus die Textgattungen dieser beiden Teilfächer beschreiben sollte. Wird TEI für andere Textgattungen (zum Beispiel Urkunden) und andere Analyseabsichten (zum Beispiel historische Forschung) eingesetzt, dann muss mit einer Schieflage bei der Anwendung ge-

rechnet werden. Ein Standard ist auch ein Wahrnehmungsfilter, mit dem die betrachteten Dinge in einem ganz bestimmten Licht erscheinen.

4.3 Anpassung und Dokumentation

Zuweilen muss beim Fehlen eines geeigneten spezialisierten Standards auf einen anderen Standard zurückgegriffen werden, der eigentlich nicht ganz genau passt. Erfolgt dann keine Anpassung an die speziellen Gegebenheiten oder aber keine Dokumentation der veränderten Benutzung, dann gerät die Hauptfunktion der Standards, nämlich berechenbare Beschreibungen zu erzeugen, in Gefahr. Man wird dann in der Weiterverwendung entweder Daten bekommen, die nicht der Sache entsprechen oder Daten, die zwar der Sache entsprechen, deren Funktionsweise aber nicht mehr dem Ausgangsstandard entspricht und deshalb nicht berechenbar ist.

4.4 Vermischung von Datenhaltung und Output

Einer der häufigsten Fehler, die auch heute noch immer wieder begegnen, ist die Verwendung von Output-Formaten für die Datenhaltung. Immer wieder gibt es Projekte, die allgemeine Daten nicht nur in PDF oder HTML darstellen, sondern diese Formate auch als zentrale Organisations- und Datenhaltungsstandards benutzen. Das ist dann legitim, wenn man entweder außer der unmittelbaren Publikation alle weiteren Verwendungsmöglichkeiten für die Daten von vornherein ausschließen kann oder wenn man der Auffassung ist, dass die zu erschließenden Objekte wirklich keine weiteren Informationen enthalten als jene, die sich auch mit den Publikationsformaten wiedergeben lassen. Wenn es zum Beispiel um Texte geht, dann ist ein Standard wie PDF als Datenhaltungsstandard nur unter der einschränkenden Prämisse vernünftig einsetzbar, dass man unter ‚Text' keine weiteren Informationen versteht als jene, die sich auf der typographischen Oberfläche abbilden lassen. Zu bedenken ist aber auch, dass Standards für die Medialisierung oft recht stark an den gegenwärtigen Hard- und Softwareumgebungen und an den gegenwärtigen Rezeptionsgewohnheiten orientiert sind. Dies steht tendenziell im Widerspruch zu einer langfristigen Nutzbarkeit, die auf möglichst allgemeinen Formen der Datencodierung und Datenstrukturierung beruhen sollte.

4.5 Mapping

Ein letztes Problem betrifft die Verwendung von Inhalten aus verschiedenen Standards. Daten müssen in der Regel gemappt werden, um sie in anderen Kontexten weiter verwenden zu können. Dabei handelt es sich um einen ganz klassischen Übersetzungsprozess, mit den bekannten Problemen der Übersetzung: Eine rein mechanische Übersetzung, die einfach von der Definition des jeweiligen Standards ausgeht, wird zwangsläufig Informationsverluste und Fehlinterpretationen nach sich ziehen. Es ist oft unumgänglich, die jeweiligen Daten und ihre Entstehungsbedingungen wenigstens auf der Projektebene etwas genauer zu verstehen, um sie sinnvoll weiter benutzen zu können. Standards bieten hier die Grundlage, um Daten überhaupt effizient ‚mappen' zu können. Sie bieten aber oft keine Garantie für das Funktionieren einer unmittelbaren, unkontrollierten und undifferenzierten Verarbeitung

5. Schluss

Der vorliegende Beitrag hat versucht, das Feld der Verwendung von Standards bei der Erschließung historischer Dokumente zu umreißen. In einem allgemeinen Modell der fortschreitenden Erschließung als einem umfassenden Informationsraum können einzelne Module isoliert werden, denen verschiedene technische Standards entsprechen. Standards können aber auch unter weiteren Aspekten beschrieben und verortet werden. Diese können innerhalb konkreter Forschungsprojekte bei der Auswahl und Anwendung geeigneter Standards ebenso helfen wie bei der Suche nach möglicherweise nicht optimalen Verwendungsweisen.