

Klaus Prator, Kreuzstr. 18 b, D-13187 Berlin
kpraetor@me.com Tel: 030-44045230

Klaus Prator

Zur Zukunft des Zitierens. Identitat, Referenz und Granularitat digitaler Dokumente¹

Zur Ontologie des Zitierens

Mit der Rede von einem Buch nehmen wir gleichermaen auf ein konkretes Exemplar wie auch auf das jeweilige Werk als Abstraktum Bezug. Die Bezugsebene ergibt sich normalerweise muhelos aus der Situation oder dem Kontext. Wo erforderlich, werden verschiedene Begriffspaare zur Unterscheidung eingesetzt, z.B. Textexemplar und Werk, Dokument (oder Textquelle) und Text. Die Inhaltsgleichheit ist kein vollig exakter Begriff. Es ist ja gerade eine Hauptaufgabe des Editors mit der Varianz innerhalb dieser Gleichheit umzugehen. Auch wenn sich ihr Spielraum nicht allgemein angeben lasst, ist es im konkreten Fall doch moglich und in vielen Fallen sogar nahe liegend und einfach, diese Gleichheit sinnvoll zu bestimmen. Bei der Entscheidung, ob zwei Dokumente als derselbe Text zu betrachten sind, kommt allerdings ein weiterer Aspekt ins Spiel, namlich die Geschichte des Textes. Bisher war ja nur von einer gleichsam zeitlosen Gleichheit die Rede. In Wirklichkeit ergibt sich diese Gleichheit oft daraus, dass ein Dokument als Drucklegung eines Manuskripts, als Abschrift oder als Neuauflage, eventuell auch als ubersetzung eines anderen betrachtet wird. Es darf nicht unterschlagen werden, dass diese Geschichte mitunter keine verburgte, sondern eine erschlossene, vielleicht gar fiktive sein kann. Die Kontinuitat des Textes geht dabei mit einer Diskontinuitat der Dokumente einher.²

Intuitiv durfte klar sein, dass das klassische wissenschaftliche Zitieren sich auf die Ebene der abstrakten Texte, der Werke, bezieht. Das inhaltliche Interesse steht im Vordergrund. Das schliet naturlich nicht aus, dass in speziellen, z.B. bibliothekswissenschaftlichen oder historischen Studien der Bezug auf Einzelexemplare erforderlich ist und insbesondere mussen naturlich im Rahmen von

¹ Dieser Beitrag bildet die Ausarbeitung eines Vortrags auf der 13. Internationalen Tagung der Arbeitsgemeinschaft fur Germanistische Edition: Medienwandel/Medienwechsel in der Editionswissenschaft vom 17.-20.2.2010 in Frankfurt

² Vgl. dazu Hans Walter Gabler, Das wissenschaftliche Edieren als Funktion der Dokumente, in: Jahrbuch fur Computerphilologie 8 (2006), S. 55-62

Editionen stets gleichermaßen Einzeldokumente wie Werke im Blick sein. Schließlich sollen referenzierbare Werke ja gerade erst konstituiert werden. Aber auch trotz der im Normalfall gegebenen Gleichgültigkeit gegenüber der materialen Beschaffenheit und auch der Medialität des Textes kann das Zitieren nicht als unabhängig davon betrachtet werden. Die Textwelt ist in der bloßen Dualität von Werk und Einzelexemplar nämlich nur sehr unzureichend beschrieben. Dazwischen existieren Auflagen, Fassungen, Übersetzungen etc., die ihrerseits nicht nur Einzelexemplare sind, aber in sich eine weiter gehende Gleichheit realisieren als die das Werk konstituierende Inhaltsgleichheit. Die Functional Requirements for Bibliographic Records (FRBR) kennen als Zwischenebenen die ‚expression‘ und die ‚manifestation‘ wobei erstere näherungsweise eine intellektuelle Ausprägung (z.B. eine Übersetzung), letztere eine physische (z.B. Druckauflage) darstellt. Derartige Zwischenebenen sind offenbar nicht universell, sondern abhängig von der jeweiligen Kultur der Schriftlichkeit, zum Beispiel verschieden in einer handschriftlichen und einer druckbasierten Dokumentenwelt.

Referenz in der Druckwelt

So springt in der Druckwelt die Bedeutung der Auflage ins Auge. Obwohl Werke in vielen Auflagen existieren können, ist hier die Auflage der eigentliche Bezugsrahmen des Zitierens. Es ist die Technik des Druckstocks (oder der Matrize), die die Gleichheit einer Ausgabe (und damit automatisch auch Inhaltsgleichheit) gewährleistet. Dementsprechend führt das klassische Zitat neben Autor und Titel Verlagsort und Jahr (der Auflage) an. Bei Mehrfachauflagen wird diese als Zahl und manchmal hilfsweise auch das Jahr der Erstauflage angegeben. Das ist natürlich deshalb von Wichtigkeit, weil neue Auflagen oft nicht unverändert erfolgen, sondern für Verbesserungen oder Veränderungen benutzt werden. Das kann zu inhaltlichen Folgen für das Zitat führen aber auch zur Veränderung des Bezugs. Üblicherweise folgt dieser ja in der Druckwelt dem Seitenfall, das heißt er erfolgt durch Angabe der Seite und eventuell auch der Zeile(n), auf die Bezug genommen wird. Da der Seitenfall sich auch bei vergleichsweise kleinen Ergänzungen oder Streichungen beträchtlich verändern kann, ist die Bezugnahme innerhalb einer Auflage problemlos, zwischen verschiedenen aber nicht zuverlässig möglich. Besonders problematisch wird das bei häufig aufgelegten Werken mit großer Tradition, so dass hier die kanonische Zitierung unter Bezug auf eine Standardausgabe (wie z.B. für Platon die Stephanus-Ausgabe von 1578) verbreitet ist. Diese kanonische Zitierung wird dann in neuen Ausgaben sozusagen als virtueller Seitenfall mitgeführt. Noch

einmal sei festgehalten: nicht das Einzelexemplar, sondern (anstelle des eigentlich intendierten Werks) die Auflage bildet die Grundlage des Zitierens; und zwar sowohl des Gesamtwerks (Autor, Titel, Ort, Jahr) wie auch der Segmentierung (Seite, Zeile). Die Gleichheit stützt sich auf die Drucktechnik. Der Bezug zwischen verschiedenen Auflagen ist nicht ohne Probleme.

In einer handschriftlichen Welt, die hier nicht näher ausgeführt werden soll, sieht das ganz anders aus: Druckstöcke existieren nicht, aber zum Beispiels Kopistenschulen als Garanten einer naturgemäß ganz anders gearteten Gleichheitsbeziehung.

Referenz im Digitalen

Wiederum anders wird die Welt des Zitierens mit dem Eintritt in die digitale Ära. Auch hier verändern sich die Gleichheitsbeziehungen innerhalb der Texte auf der Basis der sich wandelnden Technologie. Sie beruht nicht auf einer einzigen Erfindung, sondern auf einer Vielzahl technischer Innovationen bei Speichern, Bildschirmen, Prozessoren und Übertragungstechniken. Aber die Digitalisierung – die Umsetzung von Texten, Tonfolgen, Zahlen, Bildern und Computerprogrammen in binäre Werte – bildet die Klammer um alle diese Entwicklungen und ermöglicht sie damit erst in dieser Weise. Was die Zeichen angeht, so erscheint dieser Schritt doch klein im Vergleich mit dem ihn vorbereitenden und ermöglichenden der Schriftentstehung, das heißt mit der Segmentierung und Darstellung der Sprache durch einen beschränkten Vorrat diskreter Zeichen. Aber die Digitalisierung radikalisiert die Zeichen, indem sie zum einen jeden Anklang an eine grafische Form tilgt und zum anderen vollkommen von Speicher- und Ausgabemedium abstrahiert – und gerade dadurch den fließenden Übergang zwischen den verschiedenen technischen Medien ermöglicht.

Trotz der angesprochenen Gleichgültigkeit des digitalen Modells gegenüber seiner technischen Grundlage lässt sich sein Erfolg ohne diese nicht verstehen. Digitalisierung ist ja kein Vorteil an sich. Weder der antike Schreiber noch sein Leser hätten es als Fortschritt empfunden, wenn er seine Inschriften in binärer Form in Stein gehauen hätte. Vergewärtigen wir uns also zumindest beispielhaft Schritte dieser technischen Entwicklung. In einer handschriftlichen Tradition erfolgt die Herstellung eines zeichengleichen Texts durch Abschreiben. Das Erreichen dieses Ziels kann durch Unvermögen oder auch aus Absicht eingeschränkt erfolgen. Im Medium des Druckes können zeichengleiche Texte mit hoher Präzision hergestellt werden. Die Gleichheit des Druckstocks oder der Matrize ist der materielle Garant

der Zeichengleichheit. Im Computer sind Kopier-, Ausgabe- und Übertragungsvorgänge ebenfalls mit hoher Genauigkeit möglich, so dass das Ergebnis vergleichbar ist. Wo liegen die Unterschiede? Im Druckmedium ist das Buch zugleich Speicher- und Präsentationsmedium. In der digitalen Welt treten Aufbewahrung und Präsentation auseinander. Der Zusammenhang beider ‚verflüssigt‘ sich, das heißt er ist mit großer Leichtigkeit und Geschwindigkeit veränderbar. Nicht nur das. Verglichen mit dem Druck gibt es eine Vielzahl von Speicher- und Ausgabemedien mit vielfältigen Übergängen zwischen diesen. Gerade wegen dieser Vielfalt wird aber die Abstraktion notwendig. Erst durch das vereinigende Modell des digitalen Textes wird die Handhabung des Übergangs zwischen den technologisch unterschiedlichen Systemen möglich.

Selbst die leidenschaftlichsten Anhänger klassischer gedruckter Ausgaben konzedieren daher gewisse Vorteile des elektronischen Edierens, zum Beispiel die Suchmöglichkeiten, die Gelegenheit zu Verbesserungen oder Ergänzungen, ohne auf eine neue Auflage warten zu müssen, die Flexibilität, einen Text im Hinblick auf unterschiedliche Erfordernisse, Ausgabegeräte oder Leservorlieben anpassen zu können. Diese Segen des Fortschritts können aber zum Fluch oder zumindest zum Nachteil werden, betrachtet man die elektronische Publikation unter dem Aspekt der Zitierfähigkeit. Die gerade noch gelobte Flexibilität wird zur Flüchtigkeit, die die verlässliche Textgestalt, die präzise Bezugnahme, teils (im Web) sogar die pure Existenz zum Problem werden lässt.

Dabei hatte zunächst alles so gut ausgesehen: Ted Nelson³ hatte in seiner frühen universellen, aber nie so verwirklichten Bibliotheks- und Hypertextvision XANADU das endgültige Ende der Zitate angekündigt. Zitate sollten in XANADU (in dem nie gelöscht wird) stets durch direkte Verweise („Transclusions“) auf das Original ersetzt werden. Diese externe Bezugnahme wird durch ein ausgeklügeltes System von Tumblers ermöglicht, die für jedes jemals erzeugte Textdokument und auch für dessen Teile eine eindeutige und beliebig erweiterbare Referenz festlegen, in der Struktur angelehnt an die Internetadressierung. Die gegenwärtige Realität sieht anders aus: Dokumente im Web verschwinden, sind unauffindbar oder verändert. In Ermangelung eines festen Seitenfalls gehen Orientierung und Referenzierbarkeit im „Fließtext“ verloren.

³ Ted Nelson hat 1965 den Begriff Hypertext geprägt und dessen Entwicklung vielfältig gefördert und propagiert. Er entwickelte ein eigenes Hypertextsystem XANADU, benannt nach dem durch ein Gedicht von Coleridge bekannten Lustpark Kubla Khans. Wohlmeinende betrachten das World Wide Web als die Verwirklichung seiner Ideen. Er selbst meint allerdings, das WWW sei gerade das, was er mit seinen Ideen verhindern wollte.

Granularität von Texten

Die größte Ratlosigkeit herrscht wohl hinsichtlich der Frage, wie auf die Teile eines Textes Bezug genommen werden soll oder anders formuliert, wie dessen Granularität⁴ beschaffen sein soll. Die Bezugnahme beim Zitieren erfolgt in gedruckten kritischen Editionen und allgemein in wissenschaftlichen Arbeiten durch die Angabe von Seiten- und eventuell Zeilenzahl. Diese sind nun im elektronischen Medium keine unverrückbaren Bezugsgrößen mehr. Durch unterschiedliche Ausgabeformate, Schriftfonts etc. ergeben sich jeweils völlig unterschiedliche Seiten- und Zeilenumbrüche. Am deutlichsten wird dies bei der frei gestaltbaren Darstellung in einem Webbrowser oder auch beim derzeit am meisten verbreiteten eBook-Format (ePub).

Das verleitet viele Editoren zu dem *klassischen (Fehl)Schluss*:

- Korrektes Zitieren ist eine wesentliche Grundlage wissenschaftlichen Arbeitens.
- Im Druckmedium wird in wissenschaftlichen Werken (einschl. Editionen) üblicherweise nach Seite und Zeile zitiert.
- Also muss auch in digitalen Editionen ein Zitieren nach Seite und Zeile möglich sein. (Auf durch das neue Medium eröffnete Gestaltungsmöglichkeiten ist zugunsten wissenschaftlicher Strenge zu verzichten. Diese sind, sozusagen als Spielereien, allenfalls für Ausgaben mit minderen oder anderen Anforderungen zulässig.)

Dazu ist zunächst festzuhalten, dass der Anspruch an die Zitierfähigkeit natürlich in der Tat essentiell ist und ohne seine Lösung eine kritische Edition nicht sinnvoll. Mehr noch: es ist zu bedenken, dass nicht nur die pure Zitierfähigkeit einer Edition, sondern auch die Zitation im Hinblick auf eine parallele Druckausgabe und evtl. auch auf bei klassischen Autoren übliche kanonische Zitierung dabei einbezogen werden muss. Sodann aber ist energischer Widerspruch anzumelden:

- Die Beibehaltung des Zitierens auf der Basis von Seiten und Zeilen ist den Möglichkeiten des neuen Mediums NICHT angemessen.
- Sie ist KEINE für die Zukunft tragfähige Lösung.
- Genau genommen war sie NIE eine gute Lösung, sondern auch in den Zeiten des Buchdrucks ein Notbehelf.

Es gibt sicher etliche Editoren, die den durch die Orientierung an den

⁴ Zumindest in informatiknahen Bereichen ist dieser Ausdruck (in etwa: Körnigkeit) üblich für das Maß und die Art in der etwas Größeres in kleinere Einheiten zerlegt werden soll.

Druckgebräuchen erzwungenen Verzicht auf Flexibilität und Darstellungsmöglichkeiten elektronischer Publikationen für hinnehmbar halten. Aber es gibt auch unter den Anhängern der Produktion druckseitenidentischer elektronischer Ausgaben wohl keine, die bestreiten, dass Editionen auf der Basis der sachlichen Auszeichnung von Texten herausgegeben werden sollten. Textgrundlage einer Edition sollen demnach Sinn tragende Strukturen sein. Die grafische Gestaltung erfolgt erst in einem späteren Schritt. Ihnen sollte zu denken geben, dass die Orientierung an Seiten und Zeilen, gar diese zur Basis inhaltlicher Bezugnahme zu machen, eklatant gegen diese Prinzipien verstößt. Stellt die grafische Gestaltung von Überschriften und Absätzen wenigstens noch eine intentionale Gliederung dar, verdankt sich die Seiten-Zeilen-Einteilung nur dem drucktechnischen Zufall.

Das angeführte Argument hat darüber hinaus den Vorzug, konstruktiv in dem Sinne zu sein, dass es auch schon Hinweise auf seine Umsetzung gibt. Denn hinter dieser konservativen Position steht natürlich unausgesprochen das sonst eher aus der Wirtschaftspolitik bekannte TINA⁵-Argument: Es gibt keine Alternative. Aber natürlich gibt es eine, nämlich *Sinn tragende Referenz*. Statt der zufälligen Einheiten sind Sinn tragende zu wählen - also beispielsweise Paragraphen, Sätze und/oder Wörter. Dazu müssen diese natürlich nummeriert oder auf andere Weise identifizierbar gemacht werden. Das ist aber für einen Computer nicht schwieriger als das Einfügen von Zeilenzahlen durch eine Satzmaschine. Die interne Nummerierung kann je nach Verwendungszweck angezeigt werden oder nicht. Verglichen mit der Angabe von Referenzseiten/-zeilen einer kanonischen Ausgabe innerhalb einer neuen Edition ist eine satzbasierte Gliederung erst recht kein Problem. Zumindest im Grundsatz ist damit auch die Parallelität von digitalen und Druckausgaben ermöglicht, denn natürlich lassen sich im Zeitalter der elektronischen Produktion gedruckter Ausgaben diese auch mit einer Nummerierung inhaltlicher Einheiten versehen. Für eine Übergangszeit müssen Seiten- und evtl. Zeilenzähler intern in der digitalen Ausgabe mitgeführt werden..

Man kann sich hier aber nicht nur auf die Zukunft als Argument berufen, sondern auch auf die Vergangenheit. In der Antike spielte das Zitieren nach Seiten und Zeilen eine geringe Rolle - und so wichtige Dokumente einer Gesellschaft, wie die Bibel oder Gesetzestexte, wurden eigentlich immer nach Sinn tragenden Einheiten wie Paragraphen oder Kapitel und Vers zitiert. So sollten wir es künftig auch wieder halten.

⁵ Margaret Thatcher soll ihre neoliberalen Wirtschaftsreformen durchgesetzt haben mit dem Argument: There Is No Alternative (TINA).

Eisberge und Korallenriffe

Das klassische Zitieren macht bei genauerer Betrachtung eigentlich nur den sprichwörtlichen kleineren, sichtbaren Teil des Eisbergs der Referenz aus, denn diese wird zum Beispiel auch in Anspruch genommen bei Fußnoten, Querverweisen oder Kollationen. Insbesondere die editorische Arbeit besteht eigentlich in der ständigen Bezugnahme auf Texte und Textstücke in unterschiedlichen Versionen. Und dort, wo diese Arbeit durch Computer unterstützt wird, müssen auch die Programme mit dem Problem der Referenz auf Text und Textteile umgehen können. Sehr schön wird das deutlich an Peter Robinsons Projekt der ‚fluid, cooperative and distributed editions‘⁶, die er auch als Korallenriffe bezeichnet.⁷ Das wäre ein ganz anderes Thema, wenn nicht das zentrale Anliegen dieses Projekts die Einführung eines ‚Unified Text Identifiers Scheme‘ wäre. Das ist kein Zufall, denn was bei geschlossenen oder zentralisierten Editionen nach internen Regeln erfolgen kann, muss hier explizit und öffentlich gemacht werden. Bedingung für das Funktionieren größerer Arbeitszusammenhänge ist die eindeutige und verlässliche Bezugnahme auf die zu edierenden Texte einerseits und Dokumente andererseits, samt ihrer jeweiligen Teile. Tatsache und Art der Bezüge auf den unterschiedlichen Abstraktionsebenen ansprechen zu können, ist die zentrale Aufgabe des Schemas. Deshalb hat Robinson auch die Dichotomie von Dokument und Text in einer zweiten Revision des Textes noch einmal präzisierend herausgearbeitet. Dem actual text stellt er den virtual/ideal text oder auch reference text gegenüber, der in den FRBR definiert wird als „work, abstract entity; there is no single material object one can point to as the work“⁸.

Wenn ich vom Inhalt eines Buches oder eben vom Werk spreche, so geschieht das nicht als Rede von etwas Separatem, ich spreche vom Buch und signalisiere, dass es mir auf Differenzen zum Beispiel in Material und grafischer Gestaltung nicht ankommt, soweit inhaltlich sich nichts ändert. Ich signalisiere damit aber auch, dass ich nicht nur von diesem konkreten Exemplar rede, sondern zwischen inhaltsgleichen Einzeldokumenten nicht weiter unterscheiden will. Die

⁶ Peter Robinson, A Specification towards Distributed Editions, 2.1 (2007) <http://www.itsee.bham.ac.uk/DistributedEditions/UTv21.pdf>, S.1

⁷ Korallenriffe sind schöne und interessante Gebilde. Durch die Arbeit unzähliger Korallentierchen entstehen Atolle von charakteristischer und doch gleichzeitig stets individueller Form – ein Beispiel für fraktale Strukturen, wie auch Blattformen oder Baumgestalten. Sperberg-McQueen (<http://cmsmcq.com/1996/ceth9605.html>) beschrieb in seinem Trip Report auf dem Text Analysis Software Planning Meeting 1996 in Princeton metaphorisch kollaborative Editionen als Korallenriffe. Robinson knüpft an dieses Bild an.

⁸ Robinson 2007 (Anm. 6), S.2

Ununterschiedenheit beziehungsweise Ununterscheidbarkeit ist nach dem Leibnizschen Prinzip der *identitas indiscernibilium* ein Kriterium dafür, dass wir es mit einem (in diesem Fall abstrakten) Gegenstand zu tun haben. Von konkreten Gegenständen erwarten wir normalerweise darüber hinaus, dass sie Individuen sind in dem Sinn, dass sie raumzeitliche Kontinuität besitzen und mithin nicht gleichzeitig an verschiedenen Stellen existieren können. Das gilt für Werke als abstrakte oder ideale Texte nicht, und so ist auch die Bestimmung der FRBR zu verstehen, dass es keinen konkreten Gegenstand gibt, auf den als das Werk gezeigt werden kann. Das ist richtig, allerdings nur eine negative Bestimmung, denn freilich können wir erst recht nicht auf einen eigenen abstrakten Gegenstand zeigen, der das Werk darstellt. Im Licht des Gesagten ist allerdings eine andere Sichtweise möglich: Das Zeigen selbst lässt durchaus noch unbestimmt, auf welcher Gegenstandsebene wir Bezug nehmen. Wir zeigen sozusagen auf Einzelexemplar und Werk gleichzeitig. Erst die Angabe des jeweiligen Aspekts im Zusammenhang des Zeigens macht klar, welche Ebene gemeint ist.⁹

Zusammen mit geeigneter Software und Serverinfrastruktur ermöglicht dieses Referenzsystem auf der Basis Sinn tragender Einheiten beeindruckende Leistungen. Aus der langen Aufzählung¹⁰ greife ich nur drei Beispiele heraus:

➤ (für den Leser) Klassifizierung (und darauf aufbauendes Retrieval) von allen Materialien, die sich auf ein beliebiges Fragment eines beliebigen Textes beziehen, zum Beispiel: Transkripte einer speziellen Quelle für diesen Text; verschiedene Quellenmaterialien wie Digitalisierungen von Manuskripten und Druckfassungen für die Ausprägungen dieses Textes; verschiedene Textversionen (Übersetzungen, Editionen, Kollationen), Sekundärmaterial wie Kommentare und Analysen.

➤ (für den Forscher) Kollationen verschiedener Versionen, die auf den verteilten Servern der Edition vorgehalten werden, mit der Möglichkeit, sie zu speichern, zu durchsuchen, zu bearbeiten und an andere weiter zu geben.

➤ Herstellung von Verlinkungen zwischen Materialien, die sich auf ein Textobjekt beziehen: zum Beispiel Zuordnung von Bildausschnitten zu Textteilen, zum Beispiele zwischen Wörtern in dem Transkript eines Manuskripts und dem Erscheinen dieser Wörter in der entsprechenden Abbildung.

⁹ Zu diesen unterschiedlichen Paradigmen der Gegenstandskonstitution vgl. Klaus Prator, *Individuen und Referenzobjekte*. In: *Methodisches Denken im Kontext*. Hrsg. von Peter Bernhard und Volker Peckhaus. Paderborn (mentis) 2008, S. 359-370

¹⁰ Robinson 2007 (Anm. 6), S. 2 f.

Robinson gründet sein Projekt auf die Text Encoding Initiative (TEI), den einzigen verfügbaren Standard¹¹ für geisteswissenschaftliche Editionen und Textarbeit. Auch die TEI empfiehlt vorzugsweise Referenz auf der Basis Sinn tragender Einheiten, insbesondere der logischen Struktur des Markups. Dabei können vorhandene Referenzstrukturen benutzt oder auch neue aufgebaut werden, vorzugsweise unter Benutzung der Attribute xml-id und n geeigneter Tags. Aber die TEI lässt auch abweichende Zitierweisen zu bis hin zu der Orientierung an Seiten und Zeilen oder an beliebigen Textmarkierungen, beide als Milestones¹² realisierbar. Mit dieser toleranten Haltung setzt die TEI nun nicht gerade ein Fanal für die Referenz auf Sinn tragende Strukturen. Robinson geht hier einen Schritt weiter, allerdings mehr im Hinblick auf editorische Erfordernisse als für die Bedürfnisse der wissenschaftlichen Zitation. Er macht eine inhaltliche Gliederung aller Teile obligatorisch und zwar ausdrücklich bis auf Wortebene. Diese Gliederungsmöglichkeiten stellt er durch ein ausgeklügeltes System von neuen TEI-Attributen zur Verfügung, die mit allen geeigneten TEI-Elementen verwendet werden können.

So würde zum Beispiel durch folgenden Ausdruck auf das erste Wort der ersten Zeile des General Prologue der Canterbury Tales Bezug genommen

```
<div UTId="/ITSEE/work=CT">...  
<div UTId="part=GP">... <p UTId="L=1">  
<w UTId="w=1">Whan</w> ...
```

aber gleichermaßen auch durch den zusammengefassten Ausdruck

```
<w UTkey="/ITSEE/work=CT/part=GP/L=1/w=1">Whan</w>
```

Diese Technik macht Gebrauch von der XPath-Syntax. Ähnlich wie auf der Ebene von Betriebssystemen Pfade den jeweiligen Ort in der Verzeichnishierarchie bezeichnen, referieren XPath-Expressions auf die Position innerhalb der Objekthierarchie der Tags, dem so genannten DOM-Tree. Vom Wurzelement ausgehend werden alle Verzweigungen (Knoten) des Baumes bis hin zur aktuellen hintereinander geschrieben und ergeben so deren eindeutige Adressierung.

Prinzipiell wäre es denkbar, diese XPath-Ausdrücke, die die Position der Elemente in der Auszeichnungsstruktur festlegen, auch als Referenzbasis für das klassische externe Zitieren zu benutzen. Da sie aber alle die aktuelle Stelle

¹¹ Man sollte seine Standardisierungswirkung allerdings nicht überschätzen, sind doch viele Projekte unter Berufung auf ihn in einer Mischung aus TEI light und eigenem Markup entstanden, folgen also nur partiell einem Standard.

¹² Milestones werden als leere Tags realisiert. Sie können Text segmentieren oder auch paarweise klammern, stellen aber selbst keine Elemente dar und deshalb entfallen auch die üblichen Umgangsmöglichkeiten für diese, z.B die Formatierung per Stylesheet.

umfassenden Tags abbilden, werden sie schnell umfangreich, unintuitiv und schwer lesbar und sind dann für Zitierzwecke schlecht geeignet. Dieses Problem wird natürlich um so größer, je extensiver die Auszeichnung des jeweiligen Dokuments erfolgte.

Damit kommen wir zu einem grundsätzlichen Problem von XML, genauer von der Entwicklung, die es genommen hat, ein Problem, das nicht nur für Referenz und Zitation von Bedeutung ist. In den frühen Tagen von XML beziehungsweise seines Vorläufers SGML war es kein Problem, von *der* logischen Struktur eines Dokuments zu sprechen. Das Ziel war damals, die mit grafischen Mitteln ausgedrückten Strukturen von Texten, wie z.B. Überschriften, Absätze, Zitate, Listen durch deren sachliche Auszeichnung, d.h. durch ihre explizite Benennung (in spitzen Klammern) zu ersetzen. Dass es *die* sachliche Auszeichnung gebe, war damals kein Thema und musste es vielleicht auch nicht sein, markierte sie doch in der Regel die Struktur, die ein Autor selbst seinem Text gab. So konnte unterstellt werden, diese sachliche Auszeichnung repräsentiere eben *die* logische Struktur des Dokuments.

Mittlerweile geht der Gebrauch von XML aber weit darüber hinaus. Gerade in Editionen wird es verwendet für die Auszeichnung von Schreibweisen, Varianten, Autorhervorhebungen, Editoranmerkungen, diverse Registerreferenzen sowie linguistische, literaturwissenschaftliche, historische Annotationen und so fort. Unter dieser Vielzahl von Auszeichnungsinteressen lässt sich schwer an der Illusion der *einen* sachlichen Auszeichnung festhalten. In vielen Editionen erreicht das Markup ein Ausmaß, dass der eigentliche Text schwer darin zu entdecken ist. Dies ist nicht nur ein Problem menschlicher Nervenstärke. Auch Programme (zum Beispiel zur Dokumententransformation), die im Prinzip mit komplexen Strukturen besser zurecht kommen als der Mensch, müssen an sich ändernde Strukturen angepasst werden. Bei tendenziell steigender Differenzierung des Markups besteht die Gefahr, dass die prinzipiell bestehenden Umformungs- und Handhabungsmöglichkeiten wegen steigender Komplexität praktisch nicht mehr vollständig nutzbar sind.

Mit dem Schwinden des Glaubens an die eine sachliche Struktur verringert sich auch die Erwartung ihrer vermuteten Beständigkeit. Die Möglichkeit ihrer Veränderung ist aber für ihre Eignung für Referenz und Zitierung noch weit schlimmer als ihre Komplexität. Mit jedem neuen Auszeichnungselement, das in ein Dokument eingefügt wird, verändert sich ja der XPath-Ausdruck für die darin enthaltenen Textteile und somit auch die Zitatangabe. Externes Zitieren ist auf dieser Basis nicht möglich. Aber auch Editionswerkzeuge, die auf dieser Basis

arbeiten, haben ein Problem, weil sie sich auf eine interne Referenzstruktur stützen, die sie in der laufenden Arbeit verändern.

Das hier vorgeschlagene Referierungsschema geht deshalb in eine etwas andere Richtung. Nicht der gesamte XPath soll benutzt werden, sondern nur die für die Struktur des Dokuments wichtigen Elemente. In einem Sach- oder Prosatext könnten dies unterhalb der Kapitelstruktur sein: Absätze, evtl. Sätze und Wörter. Bei anderen Dokumentgattungen hätte eine sinngemäße Angleichung stattzufinden, z.B. würde man in Briefeditionen Briefe, in Gesetzen Paragraphen benutzen. Die restlichen Auszeichnungen blieben für die Referenz außer Betracht. Die Referenzstruktur bleibt zumindest innerhalb einer Textfassung fest. Die eigentlich editorischen Auszeichnungen liegen dann auf einer höheren Ebene, die nicht für die interne Referenz benutzt wird. Diese Zitierweise ist grundsätzlich für digitale und Druckwerke gleichermaßen geeignet. Darüber hinaus gibt es positive Seiteneffekte. Editionswerkzeuge (z.B. für Vokabular- und Registererstellung oder Kollation) können auf dieser Grundlage präziser und effizienter arbeiten. Insbesondere die Einbeziehung der Satzebene erschließt weitergehende Möglichkeiten für Retrieval, inhaltliche und linguistische Analyse.

Ein Problem als Lösung

Ein traditionelles XML-Problem wird damit zwar nicht geschaffen, aber verschärft: die nicht zulässige Überschneidung¹³ von Elementen. Die Auszeichnung von Sätzen oder gar von Wörtern segmentiert natürlich den Text in beträchtlichem Maße. Das ist ohnedies nicht schön, erhöht aber insbesondere die Überschneidungsgefahr mit anderem Markup. Da diese aber ohnedies durch den eben konstatierten zunehmenden Auszeichnungsbedarf steigt, lohnt es sich, über Lösungsmöglichkeiten nachzudenken. Und da es sich um ein grundsätzliches XML-Problem handelt, gibt es auch schon etliche Vorschläge dafür.¹⁴ Grob lassen sich diese in drei Gruppen einteilen: Segmentierung, Milestones und Stand-off Markup. Bei der Segmentierung wird ein überschneidendes Element so in Teile zerlegt, dass es keine Überschneidung mehr gibt. Abgesehen von der erforderlichen Arbeit ist von Nachteil, dass das Element jetzt nicht mehr als Einheit erkennbar ist. Ausgefeiltere

¹³ Bekanntlich muss in XML ein innerhalb eines übergeordneten Elements geöffnetes Element auch in diesem geschlossen werden. Andernfalls spricht man von Überschneidung (Overlapping). Die Regel bringt das zugrundeliegende objektorientierte Modell zum Ausdruck.

¹⁴ Ein guter Überblick findet sich in Steven DeRose: Markup Overlap. A Review and a Horse. In: Extreme Markup Languages 2004, (<http://conferences.idealliance.org/extreme/html/2004/DeRose01/EML2004DeRose01.html>)

Vorschläge versuchen das zu kompensieren, so z.B. MECS bzw. TexMECS von Sperberg-McQueen und Huitfeldt¹⁵ oder das Join-Element in TEI. Milestones wurden schon in einer Fußnote erläutert als leere Tags, die Text segmentieren oder auch paarweise klammern können. Stand-off Markup befindet sich außerhalb des eigentlichen Quelltextes und verweist auf Stellen oder Bereiche in ihm. Ähnlich wie Milestones, vielleicht sogar noch stärker als diese, hat es den Nachteil, dass die durch sie ausgezeichneten Elemente nicht im normalen Umfang durch die existierenden Editoren, Browser und Transformationsprogramme gehandhabt werden.

Gleichwohl ist die Idee des Stand-off Markups gerade im Zusammenhang des hier gemachten Vorschlags nicht ohne Charme. Wie das Zitieren braucht es nämlich eine verlässliche und hinreichend fein granulいたe Referenz im Text. Diese ist mit der vorgeschlagenen Absatz/Satz/Wort-Struktur eindeutig gegeben, denn genau für diesen Zweck wurde sie ja gewählt. Was nun weiteres, möglicherweise überschneidendes Markup angeht, so kann mit verschiedenen Fällen unterschiedlich umgegangen werden. Handelt es sich beispielsweise um die Notierung einer grafischen Kursivierung, so spricht wenig gegen deren Segmentierung. Die Seitenzählung einer existierenden Druckfassung lässt sich über Milestones realisieren, da sie ja für die inhaltliche Struktur des Textes nicht wichtig ist. Denkt man aber an typisch editorische Elemente wie einen Kommentar, so bietet das Stand-off Markup konzeptionelle Vorteile. Der Grundtext wird nicht nur vom Kommentar selbst, sondern auch von den Ankeren entlastet, an die dieser angebunden wird. Dies entspricht einem tradierten Ideal im Bereich der Druckeditionen: Statt der Anbringung von Anmerkungszeichen im Text erfolgt die Anbindung über die Angabe von Stellen und Bezugswörtern im Kommentar. Bei der Registererstellung, z.B. von Orts- oder Personennamen, wiederum ist es ja ohnedies der Zweck der Auszeichnung, ein externes Verzeichnis zu erstellen, dessen Elemente auf die jeweiligen Textstellen verweisen. Beides ist mit dem Stand-off Markup schon gegeben: Es hat Verzeichnisform und der exakte Textverweis ist auf Basis der schon vorhandenen Struktur ohne eigenen physischen Anker möglich. Insgesamt führt die Externalisierung von Markup zu größerer Transparenz der Dokumente aber auch der damit verknüpften Arbeitszusammenhänge. Insbesondere kann damit auch extern Bezug auf Werke genommen werden, auf die uns sonst kein Zugriff gestattet ist, und damit rücken wir - last but not least - auch wieder näher an die Vision von

¹⁵ Claus Huitfeldt, C. Michael Sperberg-McQueen: TexMECS. An experimental markup meta-language for complex documents, 2001.
(<http://decentius.aksis.uib.no/mlcd/2003/Papers/texmecs.html>)

Nelsons XANADU.

Angesichts wachsender Komplexität der Textauszeichnung wird dem Stand-off Markup in jüngster Zeit doch mehr Aufmerksamkeit zuteil – trotz der unbestreitbaren Schwierigkeiten, die seine Handhabung mit sich bringt. So sehen z.B. Sahle und Schaßan¹⁶ die Notwendigkeit des Ausbaus von Auszeichnungsmöglichkeiten für historische Zwecke, gleichzeitig aber auch die Probleme zunehmend komplexer werdenden Markups. Als Auswegmöglichkeit benennen sie in diesem Zusammenhang das Stand-off-Markup. Und Robinson, der sein Referenzierungssystem ja durch Attribute in herkömmlichen TEI-Elementen realisiert, fügt schließlich einen Appendix zu seinem zweiten Draft hinzu, wo es heißt: „while all examples presume that the attributes will be embedded in fact they could be held in separate documents. ... This would act as ‚stand-off markup‘“¹⁷

Auch die TEI unterstützt Stand-off Markup und bietet sogar einen eigenen Umsetzungsvorschlag an.¹⁸ Um den kurz zu demonstrieren, hier zunächst aus den TEI-Guidelines eine Zeile aus einem Grundtext, der in Wörter gegliedert ist,

```
<l>
  <w xml:id="w001">Scorn</w>
  <w xml:id="w002">not</w>
  <w xml:id="w003">the</w>
  <w xml:id="w004">sonnet</w>
  <w xml:id="w005">critic</w>
  <w xml:id="w006">you</w>
  <w xml:id="w007">have</w>
  <w xml:id="w008">frowned</w>,
</l>
[...]
```

und das zugehörige externe Markup.

```
<p>
  <s>
    <include
xmlns="http://www.example.org/ns/nonTEI" href="." xpointer
="range(element(w001),element(w004))"/>
    </s>
  [...]
```

¹⁶ Patrick Sahle und Torsten Schaßan: Semantisierung des Textes im Lichte und im Schatten der Text Encoding Initiative (TEI), In: KI 2009 Heft 4, S.19-24.

¹⁷ Robinson 2007 (Anm. 6), Appendix 3.

¹⁸ TEI P5: Guidelines for Electronic Text Encoding and Interchange 20.4: Stand-off Markup (<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html>)

Es hat in diesem Fall den Zweck, den Grundtext durch das Tag <s> in Sätze zu gliedern. Durch Includes werden, wie in einem Template, Passagen des Quelltexts, hier ein Bereich von 4 Wörtern, in das externe Markup eingebettet. Mittlerweile sind Includes auch als XInclude W3C-Standard. Durch sie wird aus Grundtext und externer Referenz ein per Stylesheet darstellbarer Text gemacht. Die TEI will damit den Nachteil des Stand-off Markups kompensieren, nicht direkt in einem Browser darstellbar zu sein. Sie verwischt damit allerdings den Unterschied zwischen der Darstellungsebene und der Ebene der begrifflichen Auszeichnung. Allgemeiner gefasst muss externes Markup durchaus nicht die hier vorgestellte Form haben. Es kann prinzipiell beliebiges XML-, RDF- oder auch Datenbankformat haben. Sein Platz ist auf einer dauerhaften Archivebene des Dokuments, von der aus dann die darstellbaren Sichten (Präsentationsebene) erzeugt werden.

Auf beiden Ebenen gibt es im Einzelnen eine Vielfalt von Lösungsmöglichkeiten und dementsprechend auch von Entscheidungen und Problemen. Die Fragen nach dem Wie der Umsetzung sind nämlich mit weit weniger Gewissheit zu beantworten als die grundsätzliche nach dem Ob der Sinn tragenden Referenz. So könnten auf der Archivebene die erforderlichen Zählungen auch dynamisch erfolgen statt im Markup festgehalten zu werden. Prinzipiell lassen sich alle Informationen im Stand-off auch in Attributen unterbringen ohne zu Überschneidungen zu führen. Das Markup würde ausufern, wäre aber programmtechnisch wohl leichter zu bearbeiten. Bei der Umsetzung für die Präsentationsebene werden sich die Probleme teilweise in abgewandelter Form wiederholen, beispielsweise wie dem Benutzer überlappende Markierungen dargestellt werden. Bei all dem ist auch zu bedenken, dass selbst der technisch beste Weg nicht gut ist, wenn man ihn alleine geht. Standards und Kooperationen sind also stets im Auge zu behalten.

All das wird sich selbst bei sorgfältigster Überlegung nicht im vorhinein festlegen lassen, sondern bedarf der praktischen Erprobung. Diese besteht zum einen im Experimentieren, evtl. auch Spielen, mit alternativen technischen Möglichkeiten, zum anderen aber der Einbindung in eine „ernsthafte“ Praxis, an der sich Realisierbarkeit, Tragfähigkeit und Nachhaltigkeit von Lösungen beweisen müssen. Es darf deshalb als glückliche Fügung betrachtet werden, dass sich im digitalen Teil der Würzburger Jean-Paul-Editionen¹⁹ ein Projekt mit übereinstimmenden

¹⁹ Jean Paul: Werke. Historisch-kritische Ausgabe, hrsg. von Helmut Pfotenhauer. Tübingen 2009 ff, sowie Jean Paul: Sämtliche Werke. Abteilung II: Nachlaß. Ab Bd. 9 hrsg. von

Orientierungen und Zielsetzungen gefunden hat, in dem eine erste praktische Umsetzung erprobt wird.

Den Ergebnissen soll und kann hier nicht vorgegriffen und stattdessen nur ein versöhnlicher Aspekt erwähnt werden. Für die bereits publizierten Bände stehen Dateien mit den Angaben des Seiten- und Zeilenfalls zur Verfügung. Der Kundige wird wissen, dass dieser Glücksfall alles andere als selbstverständlich ist. Zusätzlich zu der Strukturierung nach Sinn tragenden Einheiten werden diese Informationen dauerhaft mitgeführt werden. Nach Bedarf kann auf der Präsentationsebene, d.h. in den konkreten Publikationsformen, darauf zurückgegriffen und so die Referenz auf die Druckfassung per Seiten- und Zeilenzählung ermöglicht werden - auch wenn es das Ziel des Unternehmens ist, dies auf Dauer überflüssig zu machen.

Abstract

In print media scientific citation is mostly based on pages and lines. For digital media an artificial fixation of pages restricts the potential of the new medium. More than this: basing reference on a merely graphical element, not even set intentionally but by chance, is a contradiction to the principle of generalized markup. Using elements with regard to content, as paragraphs, sentences, words is a better alternative. In the paper some difficulties and some ways of an implementaton of this model are considered, among others the use of stand-off markup.