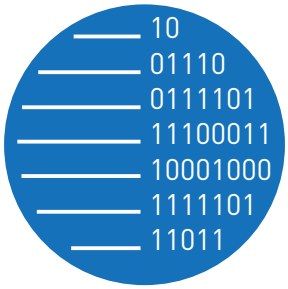

MAGAZIN FÜR DIGITALE EDITIONSWISSENSCHAFTEN

Herausgegeben vom Interdisziplinären Zentrum
für Editionswissenschaften
der Friedrich-Alexander-Universität Erlangen-Nürnberg

1 / 2015

FAU UNIVERSITY PRESS

FAU University Press
Magazin für digitale Editionswissenschaften
ISSN 2364-0855

Herausgeber:
Interdisziplinäres Zentrum für Editionswissenschaften
Prof. Dr. Boris Dreyer (Sprecher)
Universität Erlangen-Nürnberg
Department Geschichte
Alte Geschichte
Kochstr. 4, Postfach 8
D-91054 Erlangen

EDITORIAL

Das „Magazin für digitale Editionswissenschaften" versteht sich als ein offenes Forum zur Vorstellung von „best practises" für Online-Editionen. In den fünf- bis zehnseitigen Darstellungen sollen anhand konkreter Beispiele aus aktuell bearbeiteten Projekten insbesondere die „technische Seite" von Online-Editionen dargelegt werden: von der digitalien Codierung bis hin zu Visualisierungsstrategien, von theoretischen Erwägungen bis hin zu pragmatischen Überlegungen. Im Zentrum stehen fachspezifische Ansprüche, Standards und Methoden bei der Editionsarbeit sowie die verwendeten digitalen Werkzeuge und Präsentationsformen.

Die Mitglieder des Interdisziplinären Zentrums für Editionswissenschaften der Friedrich-Alexander-Universität Erlangen-Nürnberg, die das Magazin tragen, wollen mit diesem Forum einen Beitrag dazu leisten, dass Kriterien für Online-Editionen in der Öffentlichkeit zur Diskussion gestellt und die Anwendung derselben Editionswerkzeuge in verschiedenen Projekten vergleichend gegenübergestellt werden können. Auf diese Weise erfolgt eine Systematisierung der genutzten Editionsmittel und es wird möglich, gemeinsame Standards auf dem immer breiter werdenden Schnittfeld zwischen Philologie und Informationstechnik zu entwickeln.

INHALT

PSEUDO-MARKUP:
EINE ESELSBRÜCKE ZWISCHEN MANUELLER UND MASCHINELLER TEXTVERARBEITUNG

Florian Kragl

In jüngerer Zeit hat sich XML als das Mittel der Wahl durchgesetzt, wenn es darum geht, Texte gleich welcher Herkunft professionell elektronisch aufzubereiten. Sind die Texte erst einmal XML-konform gespeichert und mehr oder weniger intensiv mit Metadaten angereichert, erweist sich das Format tatsächlich als ein zuvor ungekannter Segen: Die Verbalisierung von Texteigenschaften auf allen erdenklichen Ebenen – Segmentierung, Lemmatisierung, Kommentierung, was immer – erlaubt ein rasches und präzises Navigieren durch das Korpus und dessen automatisierte Weiterverarbeitung, z. B. für linguistische oder stilistische Analysen. Wer sich allerdings je mit der Aufgabe konfrontiert sah, Textinformation in XML einzutragen, weiß, dass dieser Benutzersegen um einen Bearbeiterfluch erkauft ist. Was nämlich benutzerseitig von unschätzbarem Vorteil ist – die extensive ›Markierung‹ sämtlicher relevanter bzw. relevant erscheinender Texteigenschaften –, erweist sich bei der Dateneingabe als umständlich, zeitraubend und fehlerträchtig. Handelsübliche XML-Editoren mögen da noch eine gewisse Unterstützung bieten, weil sie einem wenigstens das Abtippen der verwinkelten Klammerstrukturen ersparen; eine rechte Hilfe sind sie aber auch nicht, gerade wenn ein Text komplex annotiert wird, weil dann aus dem Klammertippen eben ein Herumgeklicke wird, dessen Frustrationsgrad bestenfalls knapp unter dem der *plain text*-Eingabe liegt. Dieser Befund wäre im Übrigen auch projekthistorisch abzusichern: Es gibt kaum ein größeres XML-Korpus, das von Hand eingegeben worden ist; üblicherweise werden große Textmassen via Scan/OCR digitalisiert und dann erst sekundär, halbautomatisch oder von Hand, in XML weiterverarbeitet (die TextGrid-Korpora wären dafür im deutschsprachigen Raum der beste Beleg; `https://www.textgrid.de/`). Man entgeht damit der skizzierten Frustrationserfahrung; Preis dafür ist aber wiederum das Arbeiten mit ›sekundären‹ Texten, die – wenn es sich um Editionen handelt – eben *nicht* neu aus den Quellen gewonnen sind.

All dies ist intrinsische Konsequenz der XML-Struktur und betrifft ein jedes textbezogene Projekt, das sich für dieses Format entscheidet. Es gibt allerdings Bereiche, wo die damit verbundenen arbeitspragmatischen Schwierigkeiten besonders virulent werden. Dies ist immer dann der Fall, wenn (1) die XML-Code-Struktur eine Komplexität erreicht, die weit über die schiere Linearität eines Prosatextes hinausgeht, und/oder wenn (2) kei-

ne Textquellen vorliegen, die automatisch eingelesen und dann als Basis für die weitere Arbeit teilautomatisch verwendet werden können. XML-Editionsprojekte, die sich Texten aus dem Zeitalter vor der Gutenberg-Galaxis widmen, können mit beidem aufwarten: Sie gewinnen ihre Texte, gerade weil sie sich oft von alten, in die Kritik gekommenen Editionspraktiken des langen 19. Jahrhunderts absetzen wollen, neu aus Handschriften, aus Quellen also, denen mit OCR zumindest aktuell noch nicht so recht beizukommen ist: die schlicht abgeschrieben werden müssen. Und sie können aber zugleich – anders als etwa ein Editionsprojekt zu Texten der Goethe-Zeit – nicht damit Halt machen, die aus den Quellen gewonnene Textschicht im Grunde so, wie sie ist, zu präsentieren. Die Textspeicherung in mittelalterlichen oder antiken Codices ist der uns gewohnten dermaßen fremd, dass die Texte für den allergrößten Teil auch das fachnahen Publikums unbrauchbar wären, würde man nicht unablässig in sie eingreifen, um sie modernen Leseaugen erträglich zu machen: Abkürzungen müssen aufgelöst, die alte, nicht-syntaktische Interpunktion durch eine moderne ersetzt, Zeilenfall und (bei gebundener Sprache) Versgrenzen abgebildet, Korruptes verbessert (das *proof reading* ist bei Handschriften weniger streng als beim modernen Buchdruck), z. T. auch phonetisch-phonologische Disambiguierungen und Normierungen vorgenommen werden. Egal ob es nun altnordische Saga-Literatur, mittelhochdeutscher Minnesang, die gotische Wulfila-Bibel ist, die Schritte sind immer dieselben, und auf Tagungen und Kongressen wird allenthalben darüber gestöhnt, dass von modernen Projektmitarbeitern erwartet wird, was vernünftiger Weise nur von einem Computer zu fordern wäre: fließend und fehlerfrei XML, meist in der Species der TEI, schreiben und lesen zu können. Erschwert wird die Lage davon, dass die betroffenen Fächer allesamt mehr oder weniger exotisch sind, zu klein jedenfalls, um jene Geldmengen freizusetzen, deren es bedürfte, um Tools zu entwickeln, die einem wenigstens einen Teil dieser frustrierenden Handarbeit abnähmen und die Mitarbeiter aus dem *purgatorium parenthesum* erlöste (um von Handschriften-OCR gar nicht erst zu träumen).

Es war genau dieses Problem, mit dem wir uns konfrontiert sahen, als wir zu dritt – Sonja Glauch und ich in Erlangen, Manuel Braun in Stuttgart – daran machten, eine Online-Edition der mittelhochdeutschen Lyrik (Leich, Minnesang, Sangspruch) zu konzipieren. Unsere Textbasis sind Handschriften überwiegend des spätesten 13. und des 14. Jahrhunderts; diese gilt es zu transkribieren; Abkürzungen müssen aufgelöst werden; die handschriftliche Interpunktion, die meist die Strophenform anzeigt (Reimpunkte), muss dokumentiert und im Editionsprozess durch eine neuzeitliche ersetzt, umgekehrt die Strophenform durch Zeilenfall abgebildet werden; offensichtliche Fehler gehören gebessert; und schließlich wollen wir – aus Traditionsgründen, aber auch aus solchen der Didaktik – für jene Texte, für die sich dies aus sprachhistorischen und stilistischen Gründen anbietet, ›normalisierte‹ Lesefassungen herstellen, die sich von der mitunter stark dialektal gefärbten Schreibsprache der Handschriften ein Stück weit entfernen und einen

Zeichen- und Phonemsatz verwenden, den das 19. Jahrhundert als ›mittelhochdeutsche Dichtersprache‹ erfunden hat.

Dass wir unsere Daten in XML/TEI ablegen und auch anderen zur Verfügung stellen möchten, war von Anfang an klar; die konzeptuelle Aufgabe bestand darin, sie dorthin zu bringen. Wenn wir im Folgenden unseren Lösungsansatz beschreiben – dessen Idee wir im Kern Christian Aistleitner (Linz) verdanken –, so tun wir dies nicht in der Überzeugung, damit den Königsweg für das oben skizzierte Problem gefunden zu haben. Es geht uns vielmehr darum, eine Variante vorzustellen, mit der – wie wir gesehen haben – sich einigermaßen pragmatisch arbeiten lässt, und im selben Zuge zu einem Dialog über all diese kleinen Hürden anzuregen, über die nicht gerne offen gesprochen wird und mit deren Überwindung wir doch alle irgendwie auf je eigne Art kämpfen. Zu hoffen stünde, dass die Räder dann nicht immer neu erfunden werden müssten und Synergieeffekte auch jenseits der Antragsprosa tatsächlich effizient sein könnten. Da sich, soweit wir sehen, zumindest so gut wie alle Mittelalterphilologien, wenn sie sich in die Digitalität vorwagen, auf denselben Baustellen tummeln, möchte dies doch mehr sein als nur ein frommer Wunsch.

Unser erstes Prinzip ist noch keines der Dateneingabe, sondern der Datenstruktur: Wir projizieren sämtliche Textschichten, die sich aus der Art unseres Editionsprojekts ergeben, in die Linearität der Zeile. Wenn unsere Texte also beispielsweise sowohl als ›rohe‹ Transkription als auch als ›normalisierte‹ Edition vorliegen, so stehen hinter diesen beiden Ideen nicht separat abgelegte Zeichensequenzen, sondern beide werden aus ein und demselben Zeichenfluss errechnet. Dass dieser eklatant überdeterminiert ist und die Sequenzialität der Zeichenkette massiv darunter leidet, dass sie beständig und punktuell – für ein Wort, eine Silbe, häufig nur für einen Buchstaben – um zusätzliche Dimensionen erweitert wird, liegt auf der Hand; der Vorteil dieses Vorgehens aber scheint uns doch zu überwiegen, weil wir so nie Gefahr laufen, dass – etwa in Korrekturgängen – unsere Textschichten asynchron werden. Wie gesagt, dies betrifft noch nicht die Dateneingabe, sondern den Aufbau der Textdatensätze (die in unserem Projekt im Grunde Strophenbausteine sind). Und natürlich geschehen diese Vorstöße in die nächste Textschicht datentechnisch nur dort, wo die Textschichten tatsächlich auch differieren; wo ein und derselbe Buchstabe in allen Textschichten konstant ist, muss dies nicht eigens vermerkt werden.

Mit einem Beispiel: Handschriftliches *clage* wird bei uns als solches transkribiert, und es bleibt auch in der nicht ›normalisierten‹ Edition als *clage* stehen. In der ›normalisierten‹ Variante der Edition hingegen setzen wir *klage*, weil dort anlautend *c* zu *k* wird; es handelt sich um eine reine Schreibkonvention. In XML/TEI sähe dies folgendermaßen aus:

```
…
        <choice>
                <orig>
                        c
                </orig>
                <reg type=”n2”>
                        k
                </reg>
        </choice>
        lage
…
```

Wobei `<reg type=“n2“>` anzeigt, dass es sich hier um eine Normalisierung handelt (und nicht etwa um eine Textbesserung, also eine Konjektur). Man könnte auch sagen, dass wir unsere Datensätze auf diese Weise schlankestmöglich halten.

Unser zweites Prinzip beruht ebenfalls auf Effizienz, die nun aber nicht länger eine der Datenstruktur, sondern der Dateneingabe ist. Wieder aber versuchen wir, mit möglichst geringem Aufwand – heißt: mit möglichst wenigen Tastendrucken – maximale Informationsdichte zu erzeugen bzw. zu transportieren, indem mehrschichtige Informationen durch ein elaboriertes Kürzelsystem verdichtet werden. Dieses Kürzelsystem, das man als ein Sammelsurium von Tricks verstehen könnte, ließe sich am besten Pseudo-Markup nennen, und es besteht seinerseits aus einer Reihe von Komponenten, die teils auf die Zeichensatzdifferenz zwischen Handschrift und Tastatur, teils auf die Mehrdimensionalität unserer Textdatensätze reagiert:

Erstens, wofür die Maschinentastatur keine Zeichen bereitstellt, verwenden wir Kurzschreibweisen. Beispiel könnte die *a*-*e*-Ligatur sein (*æ*), die in vielen Handschriften verwendet wird, um langes umgelautetes *a* darzustellen. Wir transkribieren sie mit `#ae`. Als Operation mag dies hart an der Banalität sein; es ermöglicht es uns aber, unsere gesamte Textarbeit in *plain text* durchzuführen, weil wir für die Dateneingabe kein Zeichen benötigen, das nicht Teil des Standard-ASCII-Satzes wäre.

Zweitens, wo immer es möglich ist, komprimieren wir mehrere Textschichten in einen Eingabecode. Dies tritt besonders häufig bei Abbreviaturen auf, die in den deutschen Handschriften des Mittelalters zwar nicht so vielfältig sind wie in den lateinischen, nichtsdestotrotz aber häufig auftreten und einen nicht unerheblichen Teil des Textbestandes ausmachen. Eine der häufigsten Abbreviaturen ist der Nasalstrich, der über Vokal steht und anzeigt, dass auf diesen ein Nasal, also *m* oder *n* folgt. Wir codieren ihn, indem wir den (in der Handschrift fehlenden) Nasal ausschreiben, vor diesen aber `#` setzen. Das heißt, dass ein *e* mit Nasalstrich sowohl als `e#m` als auch als `e#n` transkribiert werden kann, je nachdem, wessen Platz der Nasalstrich vertritt. Der Textprozessor, der aus unseren Daten dann etwa eine Transkriptions- oder eine Editionsfassung (oder aber XML/

TEI-konforme Daten) herstellt, weiß, dass in beiden Fällen in der ›rohen‹ Transkription ein Strich über dem *e* zu stehen hat, in der Edition aber *em* bzw. *en*. Es ist diese Art der Komprimierung, die am Wesentlichsten dazu beiträgt, dass unsere Texteingabe zügig und – wenn man die Codes erst einmal internalisiert hat – auch einigermaßen komfortabel geschehen kann. Im Übrigen muss man nicht ein Lexikon von selbsterdachten Codes auswendig lernen, um an unserem Projekt teilzuhaben: Die Liste umfasst aktuell nicht einmal eine A4-Seite. Das liegt nicht an der Einfalt unserer Texte, sondern daran, dass wir dieses Kompressionsverfahren nur bei häufigen Textphänomenen nutzen; alles andere wäre unökonomisch.

Drittens, wo Mehrschichtigkeit nicht über komprimierte Eingaben dieser Art bewältigt wird, arbeiten wir mit abgekürzter XML/TEI-Syntax. Während es dort eine ganze Reihe von Klammern braucht, um die schlichte Information abzuspeichern, dass handschriftliches *c* in der ›normalisierten‹ Edition als *k* darzustellen ist, setzen wir:

```
{c|k}lage
```

Die geschweiften Klammern definieren den Typ der Textoperation: den einer ›Normalisierung‹, der senkrechte Strich trennt die handschriftliche Lesung links von der hergestellten rechts. Würden wir in den Text eingreifen, um eine verderbte Stelle zu bessern, geschähe dies nach demselben Muster, nur dass der senkrechte Strich nun nicht von Klammern flankiert würde, sondern von Unterstrichen. Wesentlich dabei ist, dass diese Strukturen – wie bei XML/TEI – ineinander verschachtelt sein können, etwa:

```
_ic|ich {c|k}_lage
```

Hier wird handschriftliches *iclage* zuerst zu *ich clage* gebessert, auf nächsthöherer Ebene das *c* aus *clage* in der bekannten Weise zu *k* ›normalisiert‹. In XML/TEI stünde dafür:

```
…
        <choice>
                <sic>
                        ic
                </sic>
                <corr>
                        ich
                        <choice>
                                <orig>
                                        c
                                </orig>
                                <reg type=”n2”>
                                        k
                                </reg>
                        </choice>
                </corr>
        </choice>
        lage
…
```

Hier wird in beiden Fällen – im Pseudo-Markup und in XML/TEI – mit präzise denselben Strukturen gearbeitet; Verschlankung ist hier alleine Resultat einer Verkürzung der XML-Klammernstruktur. Dass diese arbeitsökonomisch durchaus wesentlich ist, ist selbstevident.

Wichtig ist, dass das Pseudo-Markup und XML/TEI vor- und rückwärtskompatibel sind, jede Strophe also, die bereits in XML/TEI abgespeichert ist, für die Bearbeitung wieder in das Pseudo-Markup zurückverwandelt werden kann. Allerdings werden hier dann doch auch Grenzen sichtbar: In dem Moment, wo die XML-Daten mit weiteren Metainformationen – etwa solchen zur metrischen Gestalt oder einer vollständigen Lemmatisierung des Wortmaterials – angereichert wären, würde der Rücktransfer nur noch mit einem erhöhten Programmieraufwand gelingen und bliebe wohl selbst dann erheblich fehleranfällig. Was oben beschrieben ist, ist also tatsächlich primär ein Mittel der Textdatenerfassung; ob und inwieweit es später auch für Korrekturgänge taugt, wird erst noch zu prüfen sein. Seinen primären Zweck erfüllt es allerdings inzwischen schon seit etwa zwei Jahren gut, und dass wir und unsere Kooperationspartner in diesem Zeitraum und zum größeren Teil sogar ohne Drittmittelfinanzierung (diese läuft erst seit einem knappen Jahr) bereits über 2000 Strophen auf diese Weise transkribieren konnten, spricht dafür, dass so nicht nur Zeit, sondern auch Geld gespart wird.

Klar ist auch, dass dieses System keinen Universalschlüssel für Editionen ›alter‹ (vorgutenbergscher) Texte bereitstellt. Es ist zugeschnitten auf eine ganz bestimmte textuelle Situation, hier auf die deutsche Lyrik des 12. und 13. Jahrhunderts, Texte in Strophenhäppchen, in einer ganz bestimmten Sprache, überliefert mit ganz bestimmten Zeichensätzen, in sehr charakteristischen Überlieferungszusammenhängen. Im intensiven Austausch mit Kollegen benachbarter Disziplinen ist uns aber bewusst geworden, dass es sehr leicht anpassbar wäre zumindest an die Anforderungen anderer mittelalterlicher Überlieferungszusammenhänge. Heißt: Eine altnordische Saga wird andere Codes benötigen, andere Formen der Textgliederung (Prosa), ein anderes Normierungs- und Normalisierungssystem nutzen. Im Prinzip aber muss auch deren Edition auf ein mehrschichtiges Konzept – von der ›rohen‹ Transkription bis hin zum normierten oder ›normalisierten‹ Editionstext – bauen, tut dies in aller Regel auch, und darum könnte auch sie dasselbe Gerüstsystem nutzen, auch wenn die Oberflächen je nach den konkreten Gegebenheiten zu gestalten wären. Ob auch der digitalen Edition antiker Texte damit geholfen wäre, müsste man sehen; neuzeitliche Überlieferungen scheinen anders zu funktionieren, entweder stärker drucklastig, was einen Gutteil der hier beschriebenen Operationen obsolet macht, oder aber ganz strikte auf den Autograph und seine Bearbeitungsstufen konzentriert, was ebenfalls ein mehrschichtiges Editionssystem, aber eines von einer ganz anderen Art einfordert. Universal ist also der Schlüssel schon darum nicht, weil er auf handschriftliche, vielleicht noch enger: weil er auf mittelalterliche Textüberlieferung hin geschliffen ist. Wenn es mit seiner Hilfe aber gelänge, diese leichter auf- bzw. neu und

datentechnisch ambitioniert zu verschlüsseln, wäre immerhin für diesen ›mittleren‹ Bereich etwas gewonnen.

EINE DIGITALE TRANSKRIPTION DES ›DEUTSCHEN PTOLEMAEUS‹

Günther Görz / Bettina Lindner

1 Zweck des Vorhabens

Will man Aufschluss über das geographische Wissen erhalten, über das gebildete Humanisten kurz vor der Entdeckung Amerikas verfügten, so ist der sog. ›Deutsche Ptolemaeus‹ (GKW M36390, Hain 13542) eine unverzichtbare Quelle. Vermutlich im Umfeld des Nürnberger Humanistenkreises[1] entstanden und in bemerkenswert schlechter Qualität gedruckt, sind nur zwei Exemplare dieser Inkunabel erhalten. Der Text ist bis auf ein lateinisches Widmungsgedicht in einer nürnbergisch-schlesischen Varietät des Frühneuhochdeutschen verfasst, wobei der Druck eine für die Zeit nicht untypische orthographische Vielfalt aufweist. 1910 wurde von Josef Fischer in Strassburg ein kommentiertes, allerdings nicht fehlerfreies Faksimile veröffentlicht.[2] Das Exemplar der Bayerischen Staatsbibliothek wurde digitalisiert[3].

Es gibt also gute Gründe, dieses wissenschaftshistorisch interessante Werk in Form einer digitalen Transkription im Web bereitzustellen. Da es dabei auch etliche Herausforderungen zu bewältigen gibt, eignet es sich gut als praktisches Beispiel zum Erfahrungsgewinn in einem Bereich der Digital Humanities. Neben den Examensarbeiten von Barbara Ries[4] und Sarah Schulz[5] wurde ein erheblicher Teil der Arbeiten im Rahmen studentischer Übungsprojekte geleistet. Diese resultierten in zwei Web-Präsentationen der Transkription. Die erste ist eine experimentelle – aus bildrechtlichen Gründen nicht frei zugängliche – Internet-Präsentation. Weiterhin konnte mit Unterstützung des Max-Planck-Instituts für Wissenschaftsgeschichte, Berlin, die Transkription zusammen mit

1 Siehe z. B. Medizin, Jurisprudenz und Humanismus in Nürnberg um 1500. Akten der gemeinsam mit dem Verein für Geschichte der Stadt Nürnberg, dem Stadtarchiv Nürnberg und dem Bildungszentrum der Stadt Nürnberg am 10./11. November 2006 und 7./8. November 2008 in Nürnberg veranstalteten Symposien, hg. von Franz Fuchs, Wiesbaden 2010 (Pirckheimer-Jahrbuch für Renaissance- und Humanismusforschung 24).

2 Der »Deutsche Ptolemäus« aus dem Ende des XV. Jahrhunderts (um 1490), in Faksimiledruck hg. mit einer Einl. von Josef Fischer, Straßburg 1910 (Drucke und Holzschnitte des XV. und XVI. Jahrhunderts in getreuer Nachbildung). Siehe auch **https://archive.org/details/derdeutscheptole00fish**.

3 Die Seitenbilder sind einsehbar unter **http://daten.digitale-sammlungen.de/bsb00001767/image_1**.

4 Barbara Ries: Der ›Deutsche Ptolemäus‹. Vorstudien zu einer digitalen Edition, Magisterarbeit, Universität Erlangen-Nürnberg, Philosophische Fakultät II, Erlangen, Februar 2006.

5 Sarah Schulz: Zwischen Latein und Volkssprache – Der Deutsche Ptolemäus auf dem Weg zu einer vielschichtigen digitalen Edition durch die Implementierung von XLink, Bachelorarbeit, Universität Erlangen-Nürnberg, Department für Gemanistik und Komparatistik und Department Informatik, Erlangen, Juli 2010.

den gemeinfreien Seitenbildern des Fischerschen Faksimiles im Rahmen von ECHO (European Cultural Heritage Online) veröffentlicht werden[6].

2 Was ist der ›Deutsche Ptolemaeus‹?

Beim ›Deutschen Ptolemaeus‹ handelt es sich um eine kleine Kosmographie aus dem Ende des 15. Jahrhunderts. Obwohl der Autor sie als deutsche Übersetzung der ›Geographia‹ des Ptolemaeus bezeichnet, darf sie als eigenständiges Werk gelten, auch wenn sie sich erkennbar an Ptolemaeus orientiert. Besonderen Wert erhält sie durch die beigefügte, auf der Ulmer Ptolemaeus-Ausgabe basierenden Weltkarte. Eines der zwei bekannten Exemplare befindet sich in der Public Library in New York und eines in der Bayerischen Staatsbibliothek in München. Ein dritter Druck, der in Berlin gewesen sein soll, gilt seit dem zweiten Weltkrieg als verschollen.

2.1 Zum Stand der Forschung

In der Forschung ist der ›Deutsche Ptolemaeus‹ bisher eher vernachlässigt worden. Die Ergebnisse von Michael Herkenhoff[7] und Barbara Ries sind jüngeren Datums; siehe auch Brévarts Lexikonartikel[8]. Daneben erschienen bereits 1910 die von Joseph Fischer herausgegebene Faksimileausgabe[9] und bis in die 1960er Jahre hinein einige weitere grundlegende Veröffentlichungen, die sich mit unterschiedlichen Fragen zum ›Deutschen Ptolemaeus‹ beschäftigen[10, 11].

Die Frage der Autorenschaft ist nach wie vor ungeklärt. Das Buch selbst gibt keinerlei Auskunft über seinen Verfasser oder Entstehungsort. In der Forschung ist man sich seit Fischer aber einig, dass die Hervorhebung der Städte Krakau, Neisse und Nürnberg auf eine besondere Beziehung des Autors zu diesen Orten schließen lässt. So vermutet man Neisse als Geburts- und Krakau als möglichen Studienort des Verfassers.

Unsicher ist auch, in welchem Jahr der ›Deutsche Ptolemaeus‹ erschien. Da der Verfasser sowohl die 1486 erstmals gedruckte ›Peregrinatio in terram sanctam‹ des Bernhard von Breydenbach als auch die im selben Jahr von Johannes Reger besorgte Ulmer Ptole-

6 http://echo.mpiwg-berlin.mpg.de/MPIWG:X3Y43643.
7 Michael Herkenhoff: Die Darstellung außereuropäischer Welten in Drucken deutscher Offizinen des 15. Jahrhunderts, Berlin 1996.
8 Francis Brévart: Ptolemäus (›Cosmographia Phtolomei Dewtsch‹), in: Die deutsche Literatur des Mittelalters. Verfasserlexikon, 2. Aufl. hg. von Kurt Ruh u. a., Bd. 7, Berlin 1989, S. 899–902.
9 Fischer [Anm. 2].
10 Erwin Rosenthal: The German Ptolemy and its World Map, in: Bulletin of the New York Public Library 48/2 (1944), S. 135–147 plus Karte.
11 Walther Matthey: Wurde der ›Deutsche Ptolemäus‹ vor 1492 gedruckt?, in: Gutenberg-Jahrbuch 36 (1961), S. 77–87.

maeus-Ausgabe verwendete, geht man in der Forschung davon aus, dass das Werk frühestens 1486 geschrieben sein kann. Eine Abschrift des ›Deutschen Ptolemaeus‹ aus dem Nachlass Johann Schöners von 1509 markiert das späteste mögliche Erscheinungsjahr[12].

Zweifelsfrei wurde der ›Deutsche Ptolemaeus‹ mit einer Type des Nürnberger Frühdruckers Georg Stuchs gedruckt. Allerdings ergaben paläotypische Untersuchungen, dass es sich dabei nicht um die Type 14 handelte – wie Joseph Fischer noch annahm[13] –, sondern dass die kleine Kosmographie sehr viel wahrscheinlicher mit der Type 13 frühestens Mitte der 90er Jahre gedruckt wurde[14]. Für Stuchs spricht auch die Tatsache, dass dieser Ende des 15. Jahrhunderts vor allem liturgische Bücher für ostdeutsche Diözesen herausgab und zeitweise auch für das Erzbistum Krakau tätig war. Andererseits wurde in der Forschung darauf hingewiesen, dass der schlechte und ungleichmäßige Satz und die Verwendung eines spärlichen und dazu abgenutzten Typenvorrats nicht den sonstigen qualitativ hochwertigen Arbeiten Stuchs entspricht[15]. Matthey hält es daher für möglich, dass ein Anfänger im Buchgewerbe den ›Deutschen Ptolemaeus‹ mit zuvor bei Stuchs erworbenen Typen druckte[16].

2.2 Inhalt

Inhaltlich ist der ›Deutsche Ptolemaeus‹, wie die meisten Kosmographien der Zeit, zweigeteilt. Einem lateinischen Widmungsgedicht, das, so Herkenhoff, den wissenschaftlichen Charakter des Werkes betonen soll[17], folgen ein theoretisch-mathematischer erster Teil und eine Erdbeschreibung. Gerade in letzterer orientiert sich der Verfasser an der ptolemaeischen ›Geographia‹. Er beschreibt in diesem zweiten Teil alle damals bekannten Kontinente. Besonders charakteristisch für den ›Deutschen Ptolemaeus‹ ist das dezidierte Interesse an Sprachen und Alphabeten, denn der Beschreibung jedes Kontinents folgt eine Auflistung der dort gesprochenen Sprachen und Schriften. Hier ähnelt die Kosmographie zeitgenössischen Reiseberichten, »die gleichfalls ein deutliches Interesse für die Sprachen fremder Völker und Kulturen erkennen lassen«[18].

2.3 Weltkarte

Wie schon erwähnt, stellt die beigefügte Weltkarte eine Besonderheit dar. Sie ist nicht zusammen mit einer der beiden bekannten Ausgaben überliefert, sondern wurde erst

12 Cod. Vindob. lat. 2992; siehe Herkenhoff [Anm. 7]. S. 135; Matthey [Anm. 11], S. 79.
13 Fischer [Anm. 2], S. 23.
14 Matthey [Anm. 11], S. 83.
15 Matthey [Anm. 11], S. 80; Herkenhoff [Anm. 7], S. 136
16 Matthey [Anm. 11], S. 80.
17 Herkenhoff [Anm. 7], S. 139.
18 Herkenhoff [Anm. 7], S. 140.

Anfang des 20. Jahrhunderts von Joseph Fischer in der Kantonsbibliothek Vadiana in St. Gallen[19] wiederentdeckt, und zwar in einer Ulmer Ptolemaeus-Ausgabe. Sie befindet sich heute in der Public Library in New York. Dass sich weder in der gebundenen Kosmographie in New York noch in München eine Karte erhalten hat, erklärt Fischer damit, dass die Werke vermutlich nicht gebunden, sondern nur geheftet waren bzw. die Karten nur lose beilagen, um eine bequemere Nutzung zu ermöglichen[20]. Die Karte ist die erste Weltkarte, die in Planiglobular-Projektion entstanden ist und somit einen direkten Hinweis auf die Kugelgestalt der Erde gibt. Fischer konnte anhand der Übereinstimmung von Nummern in Karte und Text eindeutig die Zusammengehörigkeit feststellen[21].

3 Probleme der Digitalisierung von Inkunabeln

Die bei der Digitalisierung von Inkunabeln zu lösenden Probleme sind aus einer Reihe von Projekten hinreichend bekannt, siehe z. B. die Überlegungen von Rydberg-Cox[22] oder die Website des ›Incunable Project‹[23]. Prominent sind dabei neben der nicht normalisierten Orthographie die Vielfalt an Typen, Ligaturen, Diakritika, Abkürzungen, Wortgrenzen und -trennungen; Vieles ist dabei auch der Nähe zur spätmittelalterlichen Handschriftenkultur geschuldet. In vielen Fällen, so auch hier, ist trotz intensiver Forschung an automatischer optischer Zeichenerkennung (OCR)[24] eine automatische Transkription in maschinenlesbaren Text noch in weiter Ferne; zumindest wäre in jedem Fall der Aufwand der Nachbearbeitung im Vergleich zu einer manuellen Transkription kritisch ins Verhältnis zu setzen.

Die wesentlichen Probleme bei unserem Exemplar sind neben einem ziemlich dilettantischen Satz und den teilweise sehr abgenutzten Drucktypen die eben genannten. Interessant wäre ein Vergleich mit dem New Yorker Exemplar, denn bei Frühdrucken sind durchaus Unterschiede zwischen einzelnen Exemplaren zu beobachten; dieses konnte bisher nicht in Augenschein genommen werden und digital liegt es nicht vor. Offensichtlich stand dem Drucker ein zu geringer Typenvorrat zu Verfügung, so dass manche Typen vertauscht oder nachbearbeitet bzw. fehlende Typen aus anderen (z.B. iv für w) zusammengesetzt wurden. Die meisten der bei unserem Beispiel gewonnenen Erfahrungen sind also verallgemeinerbar – was auch eine Publikation an dieser Stelle rechtfertigen mag.

19 Siehe http://www.sg.ch/home/kultur/kantonsbibliothek.html.
20 Fischer [Anm. 2], S. 12.
21 Fischer [Anm. 2], S. 11.
22 J. A. Rydberg-Cox: Digitizing Latin Incunabula: Challenges, Methods, and Possibilities, in: Digital Humanities Quarterly 3/1 (2009), S. 1–6.
23 http://daedalus.umkc.edu/incunables/.
24 Vgl. C. Kämmerer: Vom Image zum Volltext – Möglichkeiten und Grenzen des Einsatzes von OCR beim alten Buch, in: Bibliotheksdienst 43/6 (2009), S. 626–659.

4 Ein studentisches Projekt mit TEI/XML

Am Anfang stand eine manuelle diplomatische Transkription des Münchner Exemplars als Rohtext unter Beibehaltung des Layouts und mit Ersatzdarstellungen für alle Glyphen – z. B. Diakritika wie Nasalstriche –, die im ISO-Latin-Zeichensatz nicht enthalten sind[25].

Ausgehend von dieser Transkription stellte dann Barbara Ries in ihrer Magisterarbeit systematische Überlegungen zu einer digitalen Edition an und setzte diese exemplarisch anhand des Kapitels über Europa in einer XHTML-Präsentation um. In verschiedenen Ansichten konnten in jeweils drei Spalten ein verlinkter Quellenkommentar sowie die Transkription – wahlweise diplomatisch oder bzgl. offensichtlicher Druckfehler normalisiert – und Seitenbilder (Digitalisate der BSB oder des Faksimiles von Fischer) eingesehen werden. Diese Idee auf den gesamten Textumfang anzuwenden, lag auf der Hand: Es sollte zunächst eine Web-Präsentation mit Transkription und Seitenbildern erzeugt werden, die später dann um eine neuhochdeutsche Übersetzung und um einen kritischen Kommentar erweitert werden kann.

Da grundsätzlich nur etablierte Standards und Konventionen genutzt werden sollten, gab es keine Alternative zur Codierung der Transkription in TEI/XML, womit per definitionem auch eine Entscheidung für Unicode getroffen wird. Die Umsetzung der diplomatischen Transkription nach TEI erwies sich als problemlos; Aufwand entstand durch die Auszeichnung der Druckfehler, Abkürzungen und Worttrennungen. Hinzu kamen später Auszeichnungen für "Named Entities" (Namen von Personen, Völkern, geographischen Orten) und Fachtermini. Auch wenn man ein grundsätzlich anderes Codierungs-System wählen würde, wäre der Aufwand wohl nicht geringer; letzlich hängt es von den verwendeten Werkzeugen ab, welche Belastung dem Benutzer bleibt. Der durch diese Auszeichnungen erzielte Mehrwert für Präsentation und mögliche Auswertungen ist jedenfalls enorm. Da der Text auch als Experimentierfeld für die Erprobung verschiedener Darstellungstechniken, Werkzeuge und Auswertungsverfahren dienen sollte, ist eine TEI-Codierung ein nahezu idealer Ausgangspunkt – zumal es sich ja letztlich um ein reines Textformat handelt, bei dem Nachhaltigkeit kein echtes Problem darstellt.

Auftakt des Projekts war ein Kurs bei der Sommeruniversität der Studienstiftung des Deutschen Volkes in Greifswald 2008, den der Autor zusammen mit Josef Schneeberger (Nürnberg/TH Deggendorf) mit großen Erfolg durchführte. Seitdem diente der Text immer wieder als Kern von Übungsprojekten zur mehrfach auch gemeinsam angebotenen Vorlesung ›Digitale Dokumente, Editionen und Bibliotheken‹. Sämtliche Materialien und

25 Vgl. Florian Kragl: Pseudo-Markup: Eine Eselsbrücke zwischen manueller und maschineller Textverarbeitung, in diesem Band.

Ergebnisse sind auf der Webseite ›Kulturerbe digital‹[26] und der Seite zur Vorlesung[27] zu finden.

5 Lösungsansatz

Auch wenn aufgrund der verfügbaren Ressourcen zunächst nur die digitale Publikation einer Transkription möglich war, sollten gleichwohl auch weitergehende Überlegungen zu digitalen Editionen leitend sein[28]. Durch die Wahl von TEI und Unicode waren wichtige editorische Entscheidungen vorgeprägt; darauf aufbauend wurden Transkriptionsrichtlinien erarbeitet.

Da es von unserer Inkunabel keine weiteren Ausgaben oder gar Auflagen gibt, sollte auf jeden Fall das Layout, d. h. die Zeilen- und Seitenstruktur in der Transkription erhalten bleiben. Da jedes TEI-Dokument aufgrund seiner XML-Basis aus nur einem Element mit hierarchischer Binnenstruktur besteht, können weder einander überkreuzende Hierarchien noch andere Arten von Graphen in TEI-Dokumenten direkt repräsentiert werden. Schon Seitenumbrüche durchbrechen aber die hierarchische Struktur des Dokuments, die durch die Kapitel- und Absatzorganisation gegeben ist. Daher wurden Letztere durch sog. Milestones dargestellt[29]. Für die drei hauptsächlichen Variantentypen wurde jeweils eine Darstellung mit dem choice-Element[30] vorgenommen:

— echte Druckfehler (vs. Schreibvarianten): Original und Korrektur,
— Abkürzungen: Abkürzungsglyph bzw. -form und Expansion,
— Worttrennungen (mit und ohne Trennzeichen): Getrennte und ganze Wortform (auf der Anfangszeile).

Später kamen noch Auszeichnungen von "Named Entities" hinzu, d. h. der Namen von Personen, Völkern und geographischen Orten – nach TEI P5[31] – sowie von astronomischen und geographischen Fachtermini mit dem term-Element[32].

Einige Besonderheiten, die unsere Inkunabel aufweist, sind

— eine Tabelle (der Winde)
— Marginalien (Zahlen)
— Text-Bild-Referenzen in der Form von Zahlen in Abschnitts-Überschriften, die sich in der zugehörigen Weltkarte wiederfinden.

26 http://wwwdh.cs.fau.de/IMMD8/Services/textfarm/.
27 http://wwwdh.cs.fau.de/IMMD8/Lectures/DIGIDOK/.
28 Siehe Patrick Sahle: Digitales Archiv und Digitale Edition. Anmerkungen zur Begriffsklärung, in Literatur und Literaturwissenschaft auf dem Weg zu den neuen Medien. Eine Standortbestimmung, Zürich 2007, S. 64–84.
29 Lou Burnard, Syd Bauman: TEI P5: Guidelines for Electronic Text Encoding and Interchange, Handbook, TEI Consortium, Oxford u. a. 2008, Kap. 3.10.
30 Burnard [Anm. 29], Kap. 3.4.
31 Burnard [Anm. 29], Kap.13; siehe auch http://www.teic.org/release/doc/tei-p5-doc/en/html/ND.html.
32 Burnard [Anm. 29], Kap.3.

Einige handschriftliche Marginalien wurden bisher nicht beachtet, sind aber sicher für die Erforschung der Benutzung des Exemplars wichtig. Allerdings wurde es beschnitten, so dass einige Marginalien nur fragmentarisch erhalten sind. Interessant wäre auch ein Handschriftenvergleich mit den beiden Abschriften – mit der schon erwähnten von Johann Schöner 1509 (heute in Wien) sowie einer aus dem Besitz von Siegmund Scheufler (Anf. 16. Jh., BSB München, Clm 388).

Für die TEI-Codierung des ›Deutschen Ptolemaeus‹ mit allen genannten Merkmalen wurden verschiedene Editoren eingesetzt, die alle zumindest über Module für die Syntax von XML verfügen. Von großem Vorteil ist, wenn beim Anlegen eines Elements automatisch die schließende Klammer erzeugt wird und wenn Schablonen für komplexe Konstrukte wie das choice-Element definiert werden können oder gar automatisch angegeben werden. Für die vereinfachte Texteingabe bietet sich auch an, wie Kragl[33] beschreibt, kurze Ersatzdarstellungen zu verwenden, die dann mittels eines Texteditors, z. B. sed (Unix) im Batch-Modus, in die entsprechend instantiierten Schablonen expandiert werden.

Zunächst wurde – und wird – Emacs[34] benutzt, dessen Benutzerfreundlichkeit allerdings nicht alle so positiv einschätzen wie der Autor. Als weitere kostenfreie Alternative wurde das XML-Modul für die Java-orientierte Programmentwicklungsumgebung Eclipse[35] eingesetzt. Sofern man sich den Rahmenvorgaben beugt, nämlich ein Editionsprojekt im Stil eines Software-Entwicklungsprojekts zu organisieren, ist das XML-Plugin zweckdienlich. Zusätzlich hat man den Vorteil, dass mit Einrichtung einer geeigneten Speicherorganisation die Daten in einem Mehrbenutzermodus in verteilten Rollen bearbeitet werden können. Dies erspart den Einsatz einer sog. Groupware, wie wir sie am Anfang des Projekts eingesetzt hatten[36]. Seine Benutzung ist inzwischen einfacher geworden im Zusammenhang mit einem anderen XML-Editorprojekt, ArborealMWN[37]. Arboreal war ein Projekt von Harvard mit dem MPI für Wissenschaftsgeschichte, das vor allem die Integration linguistischer Dienste – Lexika, Lemmatisierung und Morphologie, Terminologieverwaltung – bei der Bearbeitung digitaler Editionen unterstützt. Die aktuelle Implementierung auf der Basis von Eclipse, die die Handhabung wesentlich vereinfacht und die weitgehend fertiggestellt ist, hat das o. g. XML-Modul integriert. Als dritter Editor wurde Oxygen[38] eingesetzt, der kostenpflichtig, aber äußerst komfortabel ist, jedoch eine solide Kenntnis von XML voraussetzt – und deshalb für Bearbeiter mit einem philologischen Hintergrund eine gewisse Hürde darstellen dürfte. Angeblich können spezielle

33 Kragl [Anm. 25].
34 http://www.gnu.org/software/emacs/.
35 https://www.eclipse.org/.
36 Konkret http://bscw.de von Fraunhofer, freie ›educational license‹.
37 http://sourceforge.net/projects/arboreal/.
38 http://www.oxygenxml.com/.

Konfigurationen für bestimmte Aufgaben generiert werden; hierzu liegen bisher keine Erfahrungen vor.

Ist die Codierung des Textes in TEI einschließlich der Metadaten im TEI-Header abgeschlossen, steht die Frage der Präsentation an. Hierzu gibt es ein breites Spektrum an Möglichkeiten, wobei die einfachste darin besteht, die TEI-Datei nach XHTML zu konvertieren, so dass sie mit jedem Browser gelesen werden kann. Allerdings ist bei der Verwendung von Elementen, die, wie z. B. choice, Redundanzen einführen, ein Vorverarbeitungsschritt notwendig, um zu verhindern, dass z. B. eine Abkürzung als solche und daneben auch noch in expandierter Form wiedergegeben wird. Beispielhafte XSLT-Skripte sind auf der Webseite zur o.g. Vorlesung zu finden. Ihre Anwendung kann entweder im Kommandozeilenmodus durch expliziten Aufruf eines Prozessors wie Saxon erfolgen oder komfortabler eingebunden in einen XML-Editor – optimal mit Oxygen. Eine einfache Möglichkeit zur Konvertierung (und viele andere!) ist über die Website **http://www.tei-c.org/oxgarage/** verfügbar. Möchte man die XHTML-Transkription zusammen mit den entsprechenden Seitenbildern anzeigen, kann dies – wie bei Ries – in zwei Frames erfolgen, wobei die Verknüpfung in der TEI-Datei z. B. durch das facsimile-Element[39] angegeben werden kann.

Anstelle einer Transformation nach XHTML bietet ›TEI Boilerplate‹[40] die Alternative der direkten Anwendung eines Stylesheets auf die TEI-Datei. Im einfachsten Fall muss nur eine Operationsanweisung in die TEI-Datei eingefügt werden, die auf ein geeignetes Stylesheet verweist, wobei eine Reihe von Stylesheets mitgeliefert wird. Voraussetzung ist allerdings, dass ein Apache Webserver installiert ist[41] und die teibp-Dateien dort auch vorgehalten werden. Auch hier ist bei Redundanzen eine vorherige Transformation erforderlich; in einfachen Fällen kann die Auswahl aber auch über ein im Stylesheet implementiertes Toolbox-Element getroffen werden.

6 Aufbereitung für eine Web-Präsentation

Zur Web-Präsentation gibt es, wie oben erwähnt, zwei Varianten. Beide erlauben, durch Schaltflächen die Art der Normalisierung einzustellen. In der ersten, experimentellen Variante wurde auch Javascript-Code implementiert, durch den ermöglicht wird, bei Positionierung des Mauszeigers im Bild die entsprechende Zeile in der Transkription durch Einrahmung hervorzuheben. Sie benutzt allerdings lokal bearbeitete und gespeicherte Seitenbilder des BSB-Digitalisats. Die BSB lässt aber keine Gesamtpräsentation ihres Digitali-

39 Burnard [Anm. 29], Kap. 11.1.
40 **http://dcl.slis.indiana.edu/teibp/index.html**.
41 Am einfachsten mit dem PHP-Entwicklungspaket XAMPP, **https://www.apachefriends.org/de/index.html**.

sats zu, sondern nur die Verlinkung von einzelnen Bildern. Die ECHO-Version benutzt aus diesem Grund nur die Bilder des Fischer-Faksimiles.

In ECHO sind verschiedene Anzeigevarianten vorgesehen, u. a. Mini-Seitenbilder (Thumbnails) als Gesamtübersicht oder zusammen mit der Transkription – in verschiedenen einstellbaren Normalisierungen – oder XML-Code oder Vollbild. Der XML-Code kann auch im Ganzen heruntergeladen werden. Intern steht eine Reihe spezieller Werkzeuge zur Verfügung, u. a. zur Einfügung von Satzgrenzen-Auszeichnungen. Dies wurde von den Administratoren auch angewandt, so dass die angebotene Volltextsuche jeweils die Treffer im Kontext des jeweiligen Satzes anzeigt. Nebenbei bemerkt waren auch hier wegen der nicht normalisierten Zeichensetzung manuelle Eingriffe notwendig. Wünschenswert wäre noch eine Anzeige von Paralleltexten, etwa des Textes in Ausrichtung (alignment) mit einer Übersetzung.

Weiterhin gestattet ECHO auch, jedes Wort mit einem Lexikoneintrag zu verknüpfen – sofern entsprechende Lexika zur Verfügung stehen. Dies trifft zum einen auf das lateinische Widmungsgedicht zu, wo u. a. eine Verlinkung zum (gemeinfreien, da 19. Jh.!) Lexikon von Lewis und Short und darüber dann zur Perseus Digital Library[42] besteht. Für unsere Varietät des Frühneuhochdeutschen gibt es kein geeignetes Lexikon, so dass nur auch im modernen Deutsch gültige Wortformen verlinkt werden. Das führt wegen der durch die nicht normalisierte Orthographie bedingten Ambiguitäten bisweilen zu bizarren Ergebnissen, etwa im Beispiel »schonen figuren«, wo zwar »schönen« gemeint ist, aber auf der (vom Benutzer konfigurierbaren) Lexikonseite dann Zugriffe für »schonen« in digitalen Lexika des Neuhochdeutschen (u. a. DWDS mit Korpusbelegen, Wortschatz Leipzig inklusive Dornseiff-Thesaurus, auch CELEX-Morphologie) angezeigt werden.

Für die interne Webpräsentation wurde auch die native XML-Datenbank eXist-db[43] zur Speicherung der TEI-Datei erprobt. Die Handhabung hat sich als recht einfach erwiesen, bringt aber in unserem Fall einer einzigen TEI-Datei keine Vorteile; solche sind aber bei größeren Korpora zu erwarten.

7 Verarbeitung

Für die Untersuchung der sprachlichen Form des Textes bieten sich aufgrund der sehr eigenen Orthographie im Prinzip nur Werkzeuge an, die Wortformen als bloße Zeichenketten ohne Rückgriff auf linguistische Informationen verarbeiten. Dies ist mitnichten ein Sonderfall, denn dasselbe gilt für viele seltene Sprachen; es gibt schlechterdings kaum Lexika und Sprachmodelle für Varietäten mit nichtnormierter Orthographie. So sind auch für

42 http://www.perseus.tufts.edu/.
43 http://exist-db.org/.

moderne Sprachen erzeugte Tagger, z. B. TreeTagger[44], bestenfalls von heuristischem Wert. Selbst für alte Sprachen wie Griechisch und Latein gibt es bisher kaum linguistische Software, die über die morphosyntaktische Verarbeitung hinaus auf die Ebene der syntaktisch-semantischen Strukturen reicht.

Wir müssen uns also auf Wort(formen)listen – auch rückläufige für die Suffixanalyse – mit Häufigkeiten, Bigramme, Trigramme, Konkordanzen und eine Vollformenindexierung beschränken. Am einfachsten geht dies mit Rohtexten ohne Auszeichnungen; hierfür stehen in der Unix-Welt diverse über die Kommandozeile aufrufbare Programme zur Verfügung. Unter den einschlägigen integrierten Programmpaketen und Webdiensten gibt es inzwischen etliche, die auch mit XML-Tags – abgesehen von Redundanzen – umgehen können, u. a. haben wir mit dem Korpus-Analysewerkzeug antconc[45] sehr gute Erfahrungen gemacht. Unter den webbasierten Korpus-Tools ist Voyant[46] besonders zu empfehlen.

Auf der Basis erzeugter Wortformenlisten wurden Untersuchungen durchgeführt, inwieweit unscharfe Suche in Texten mit nichtnormierter Orthographie sinnvoll eingesetzt werden kann. Zugrunde gelegt wurde die sog. Edit-Distanz (Levenshtein-Algorithmus[47]) zwischen zwei Zeichenketten. Sie wird auch in dem Unix-Programm ›agrep‹ verwendet, das zwar prinzipiell geeignet erscheint, aber noch genauer evaluiert werden muss.

Für weitergehende linguistische Analysen bietet sich als Notbehelf an, mit textnahen, d. h. eher wörtlichen Übersetzungen als Paralleltext zu arbeiten. Aus einem anderen Projekt liegen gute Erfahrungen mit dem Stanford-Parser[48] vor. Während dessen deutsches Sprachmodell nur Konstituentenstrukturen erzeugen kann, liefert das englische auch Dependenzstrukturen, die für die semantische Ebene besser anschlussfähig sind. Zur Erzeugung von Prädikat-Argumentstrukturen hat sich (für das Englische) u. a. der Semantic Role Labeler[49] der Universtität Lund bewährt, der auch den Zugriff auf ein framebasiertes Lexikon ermöglicht.

8 Erstellung von Druckfassungen

Zusätzlich zur Online-Präsentation wird oft auch eine Ausgabe in einem Druckformat, normalerweise PDF, oder in einem anderen eBook-Format gewünscht. In der XML-Welt wurde hierzu die Transformation mittels XML-FO (Formatting Objects) entwickelt. Da zwischen der inhaltlich-deklarativen Auszeichnung in TEI/XML und einer Auszeichnung

44 http://www.cis.uni-muenchen.de/schmid/tools/TreeTagger/.
45 http://www.antlab.sci.waseda.ac.jp/software.html.
46 http://voyant-tools.org/.
47 Siehe http://www.levenshtein.de/.
48 http://nlp.stanford.edu/software/lex-parser.shtml.
49 http://nlp.cs.lth.se/.

mit LaTeX[50] eine Analogie besteht, liegt die Vermutung nahe, dass zur professionellen Layout-Produktion mittels XML-FO ein ähnliches Verfahren einzusetzen wäre. Tatsächlich gab es auch Versuche, die Satzmaschine von TeX für diesen Zweck einzusetzen, die allerdings bisher aus nicht klar nachvollziehbaren Gründen nicht besonders erfolgreich verliefen. Für diesen Zweck eigens entwickelte Open-Source–Tools hatten bisher nur einen begrenzten Leistungsumfang. Seit kurzem steht mit XML-Print[51], das im Rahmen des Verbundprojekts TextGrid entwickelt wird, ein neues, ebenfalls mit Eclipse entwickeltes Rahmensystem mit einer flexiblen graphischen Benutzungsschnittstelle bereit. Erste Versuche mit unserem Text verliefen erfolgreich, jedoch ist noch einiger Feinschliff bei den Ausgabeformaten notwendig. In der Handhabung besteht gegenüber dem reinen XML-FO-Konzept eine signifikante Vereinfachung.

9 Ausblick

Mit der TEI-Codierung ist eine Basis für eine Reihe künftiger Erweiterungen gelegt. In einer Bachelorarbeit[52] wurde u. a. untersucht, inwieweit sich die XML Linking Language (XLink) zur Definition und Nutzung von Links in TEI-Dokumenten einsetzen lässt. Zwar verfügt die TEI auch über eine eigene Link-Komponente, aber die Nutzung der damit formulierten Links wird bisher nicht von Werkzeugen unterstützt. Auch XLink wird eher stiefmütterlich behandelt; doch es konnte praktisch nachgewiesen werden, wie man durch eine Auswertung von XLink-Konstrukten zu einer erweiterten HTML-Edition gelangt.

Ein großes Potential besteht in der semantischen Auswertung der Auszeichnungen von Personennamen, Toponymen, Zeitangaben, Ereignissen etc. durch deren Anschluss an eine semantische Modellierung durch formale Ontologien wie CIDOC CRM[53]. Das Conceptual Reference Model (CRM) wurde als Referenzontologie für die Dokumentation des Kulturerbes definiert und ist ein internationaler Standard (ISO 21127). Eines seiner wesentlichen Ziele ist die Herstellung von Interoperabilität und umfassender Vernetzung unterschiedlicher Ressourcen. Die Überlegung »that detailed information about persons (physical and legal), dates, events, places, objects etc. and their interpretation could be marked up outside the text, and that this could be connected to on-going ontology work« führte zur Gründung einer Ontologie-Arbeitsgruppe innerhalb der TEI[54]. Durch die Einbindung von Normdateien, z. B. für geographische Bezeichnungen, die wie

50 http://www.latex-project.org/.
51 http://kompetenzzentrum.uni-trier.de/de/projekte/projekte/xml-print/.
52 Schulz [Anm. 5].
53 http://www.cidoc-crm.org/, http://erlangen-crm.org.
54 Siehe http://www.tei-c.org/Activities/SIG/Ontologies/.

Pleiades eindeutige URIs zuordnen[55], wird eine Vielfalt weiterer Ressourcen zugänglich. Eine interssante Möglichkeit besteht in der Geovisualisierung mit Web-GIS–Systemen wie GeoTemCo[56].

Zugleich können die aus den Auszeichnungen gewonnenen Propositionen in der Subjekt-Prädikat-Objekt-Form, sog. RDF-Tripel[57], in Graphen organisiert als »Linked Open Data« über eine standardisierte Anfrageschnittstelle (SPARQL[58]) allgemein zugänglich gemacht werden. Beispielhaft sei hier die von Thomas Kirchner et al. besorgte digitale Edition von Joachim von Sandrarts ›Teutscher Academie der Bau-, Bild- und Mahlerey-Künste‹, Nürnberg 1675/1679/1680, genannt[59]. In diesem Editionsprojekt wird besonders darauf hingewiesen, »dass es sowohl machbar als auch wünschenswert erscheint, zumindest einen Teil unserer Daten als Linked Open Data zu veröffentlichen: Personen, Orte, Kunstwerke, Bibliographie-Objekte und die Bezüge zwischen diesen.« Durch diese neue Dimension der Interoperabilität gewinnt nicht nur die eigene Edition einen erheblichen Mehrwert, sondern dieser kommt in gleicher Weise auch anderen zu.

Danksagung

Besonderer Dank gebührt Josef Schneeberger und Martin Scholz für die umfangreiche und kontinuierliche technische Untersützung des Projekts. Ohne die vielen Beiträge von Studierenden zu Transkription, Korrekturlesen und Implementation – neben den bereits Genannten die Teilnehmer unseres Kurses bei der Sommeruniversität der Studienstiftung des Deutschen Volkes in Greifswald 2008 sowie in Erlangen Studierende aus Buchwissenschaft, Germanistik und Informatik – wäre das Projekt immer noch in den Kinderschuhen. Für die Integration in ECHO am MPIWG Berlin sei Josef Willenborg und Klaus Thoden herzlich gedankt.

55 Siehe Pelagios, http://pelagios-project.blogspot.de/, dort speziell auch IN USE.
56 http://www.informatik.uni-leipzig.de:8080/geotemco/.
57 Siehe http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.
58 http://www.w3.org/TR/rdf-sparql-query/.
59 http://ta.sandrart.net/de/, insbesondere http://ta.sandrart.net/de/info/services/lod-rdf/.

Und Semantik wuchs in ›EDEN‹
– eine Vorstellung und ein Erfahrungsbericht

Martin Scholz / Marvin Holdenried / Boris Dreyer /
Klaus Meyer-Wegener / Günther Görz

1 Einleitung

Die ›Epigraphische Datenbank Erlangen-Nürnberg (EDEN)‹[1] ist eine Online-Datenbank für antike Inschriften aus Kleinasien. Momentan beschränkt sie sich auf die drei antiken griechischen Siedlungen Metropolis in Ionien, Magnesia am Mäander und Apollonia am Rhyndakos.[2] Die Datenbank wird seit 2012 an der FAU unter Federführung von Boris Dreyer (Professur für Alte Geschichte) sowie der AG Digital Humanities aufgebaut.

Technisch basiert EDEN auf der virtuellen Forschungsumgebung ›WissKI‹[3], welche von der AG Digital Humanities in Zusammenarbeit mit dem Germanischen Nationalmuseum (Nürnberg) und dem Zoologischen Forschungsmuseum Alexander Koenig (Bonn) entwickelt und gepflegt wird. Das WissKI-Projekt und die daraus hervorgegangene Software entstanden aus der Unzufriedenheit mit Datenmanagement-Werkzeugen für das Kulturerbe und dem Wunsch, diese zu beheben.

Ein besonderes Augenmerk wurde dabei dem Problem der schnell alternden Datensilos[4] gewidmet. Der Alterungsprozess wird verhindert – so die grundlegende Idee –, wenn das Datensilo stets in den wissenschaftlichen Prozess eingebunden ist, neue Erkenntnisse dorthin zurückfließen und die darin abgelegten Daten über die Projektlaufzeit hinaus nutzbar bleiben. Daher hat WissKI den Anspruch, verschiedene Aspekte wis-

1 Erreichbar unter ›wisski.cs.fau.de/eden‹; eine Einführung zu Online-Datenbanken im allgemeinen und zu EDEN im speziellen in Marvin Holdenried, Charlotte Roueché, Martin Scholz: Digital Epigraphy in archaeological context: The case of Metropolis, Magnesia & Apollonia, in: Die Surveys im Hermos- und Kaystrostal und die Grabungen an den Thermen von Metropolis (Ionien) sowie am Stadion von Magnesia am Mäander. Neue Methoden und Ergebnisse, hg. von Boris Dreyer, Berlin, Münster 2014 (Orient & Okzident in der Antike 1), S. 163–183.

2 Siehe hierzu allgemein Dreyer, Die Surveys [Anm. 1].

3 Siehe ›wiss-ki.eu‹ sowie Martin Scholz, Guenther Goerz: WissKI: A Virtual Research Environment for Cultural Heritage, in: European Conference on Artificial Intelligence (ECAI) 2012, hg. von Luc De Raedt et al., Montpellier 2012, S. 1017f.

4 Damit sind Datenbestände gemeint, die z. B. aus organisatorischen oder Workflow-Gründen nicht mehr gepflegt werden, sodass mit fortschreitender Zeit Aktualität und/oder Interpretierbarkeit der Daten stark abnehmen (Alterung).

senschaftlichen Arbeitens in einem System zu vereinen. Insbesondere greifen Eingabe, Bearbeitung und (Online-)Publikation der Daten auf denselben Datenbestand zu.

In den folgenden Abschnitten werden einige Vorteile von WissKI für die Umsetzung der Epigraphischen Datenbank erläutert. Der Artikel klingt aus mit einer kurzen, allgemeineren Diskussion, wie der Einsatz der WissKI-Software möglichst optimal vorbereitet werden kann.

2 Webauftritt

WissKI ist ein Content Management System und als solches eine Software zur Verwaltung von Webauftritten. Es ist also genuin web-basiert und eine mit WissKI umgesetzte Forschungsdatenbank wie EDEN ist daher immer auch ein Webauftritt; eventuell passwortgeschützt. Obwohl EDEN primär für die Online-Publikation gedacht ist, können die Freitexte und tabellarischen Daten in EDEN auch zur Vorbereitung einer neuen Printedition über die Inschriften von Metropolis, Magnesia und Apollonia genutzt werden.

Wie bei Wikipedia erfolgen Eingabe, Bearbeitung und Anzeige von Daten lediglich über das Internet, typischerweise im Browser. Die Daten liegen nicht auf einem oder mehreren PCs, sondern zentral auf dem Server, auf dem WissKI eingerichtet ist. Dies birgt Vorteile hinsichtlich Synchronisation und Sicherung der Daten. Neben der Verwaltung der forschungsrelevanten Daten gibt es umfangreiche Funktionalität zum Aufbau der eigenen Webpräsenz, vom Erstellen einfacher Webseiten (z. B. Startseite oder ›Über uns‹) und eines eigenen ›Look & Feel‹ über Kommentarfunktionen bis zu einer feingranularen Benutzer- und Rechteverwaltung. So präsentiert EDEN dem Besucher etwa eine Einstiegsseite mit Projektbeschreibung, während die Entwickler als geschlossene Gruppe in Projekt-TODO-Listen auf der Plattform die Prioritäten der weiteren Entwicklung diskutieren können.

3 Integration verschiedener Disziplinen

Inhaltlich wird EDEN von den Erlanger Althistorikern in enger Zusammenarbeit mit Archäologen der Universität Erlangen-Nürnberg sowie den archäologischen Kollegen vor Ort in der Türkei entwickelt. Ziel ist, die teilweise sehr unterschiedlichen Perspektiven der beiden Disziplinen in einer Datenbank zu vereinen. Denn während sich Althistoriker primär mit den immateriellen Eigenschaften der Inschrift, also Textinhalt und -form, beschäftigen, ist für Archäologen der materielle Träger der Inschrift von größerem Interesse. In bestehenden Datenbanken und Publikationen findet sich je nach Disziplin die eine

oder andere Seite vernachlässigt.[5] Durch die enge Kooperation und die Möglichkeiten digitaler Medien wird mit EDEN eine Brücke zwischen den beiden Disziplinen geschlagen. Da die Datenbank stetig erweitert werden kann, können inhaltliche Informationen sowie Informationen zum physischen Inschriftenträger in einem Detailgrad erfasst und kombiniert werden, der den Ansprüchen von Althistorikern wie auch Archäologen genügt. Dabei ist der Aufbau der Datenbank nicht auf diese beiden Disziplinen begrenzt. Entsprechend den Entwicklungen durch Kooperationen und neue Forschungsschwerpunkte können ehemalige ›Randinformationen‹ schnell in den Fokus der Datenbank rücken. Das zugrundeliegende System macht einen solchen Fokuswechsel leicht, da Daten nicht neu erfasst werden müssen, sondern lediglich die bereits bestehenden ›Randinformationen‹ angereichert werden und somit der Detailgrad auf die neuen Anforderungen angehoben werden kann. So könnten die bereits vorhandenen geographischen Informationen durch eine Kooperation mit Geologen ausgeweitet werden und EDEN zu einer für (kultur-)geographische Forschungen nutzbaren Quelle machen.[6]

4 Wissensrepräsentation: Datensätze und Wissensnetz

Diese Flexibilität beruht wesentlich auf den angewandten Techniken zur Wissensrepräsentation aus dem Umfeld des sogenannten ›Semantic Web‹: Dazu wird anstatt einer relationalen Datenbank eine Graphdatenbank, ein sogenannter *RDF* Triple Store, eingesetzt, der Informationen in Form von Netzwerken anstatt Tabellen repräsentiert. Durch Ontologien werden die eingegebenen Informationen weiter strukturiert und kategorisiert und für den Computer semantisch greifbar gemacht. Gleichzeitig bleibt die Struktur durch die Netzform flexibel und die Datenbank offen für neue Strukturen. Dieser Ansatz ist etwa mit dem Erstellen von Mindmaps und ähnlichen Brainstorming-Methoden vergleichbar. Ein wesentliches Merkmal dieser semantischen Netze ist die Auflösung beziehungsweise Relativierung des Datensatzes. Anstatt Wissen aufgeteilt in Datensätzen mit mehr oder minder festgefügter Struktur zu verwalten, stellen semantische Netze Wissen als Netz aus Kanten und Knoten dar. Dabei repräsentieren typischerweise die Knoten Objekte des Anwendungsbereichs und Kanten Relationen zwischen diesen Objekten. Änderungen beim Detailgrad des zu erfassenden Wissens lassen sich jederzeit leicht umsetzen, indem neue Knoten und Kanten hinzugefügt werden. Abbildung 2 zeigt die Informationen aus Abbildung 1 als Netz. Die Darstellung entspricht jedoch nicht exakt der in EDEN, sondern sie wurde zu Illustrationszwecken vereinfacht. Man kann gut erkennen, dass bestimmte Metadaten enger miteinander in Verbindung stehen als andere.

5 Gründe dafür können vielfältig sein: Platzbeschränkung bei Printmedien mögen ebenso eine Rolle spielen wie auch Desinteresse oder mangelnde Kenntnisse der jeweils anderen Disziplin.

6 Siehe hierzu Mark Vetter: Geodätische Erfassung und GIS-gestützte Darstellung der antiken Stadt Metropolis/Ionien, in: Dreyer, Die Surveys [Anm. 1], S. 125–136.

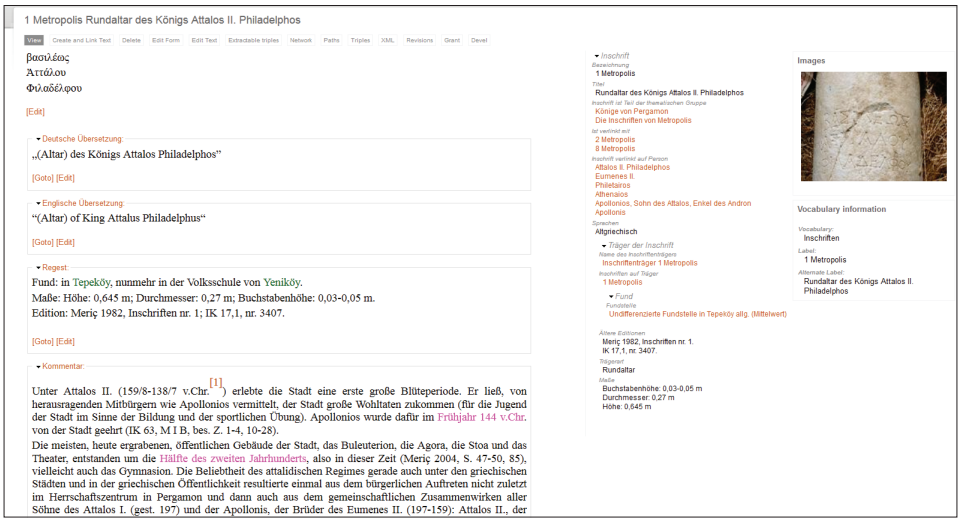

Abbildung 1: EDEN: Inschrift 1 Metropolis, Rundaltar des Königs Attalos II. Philadelphos, URL: **http://wisski.cs.fau.de/eden/content/ecrm_E34_Inscription01e6**, Standardansicht

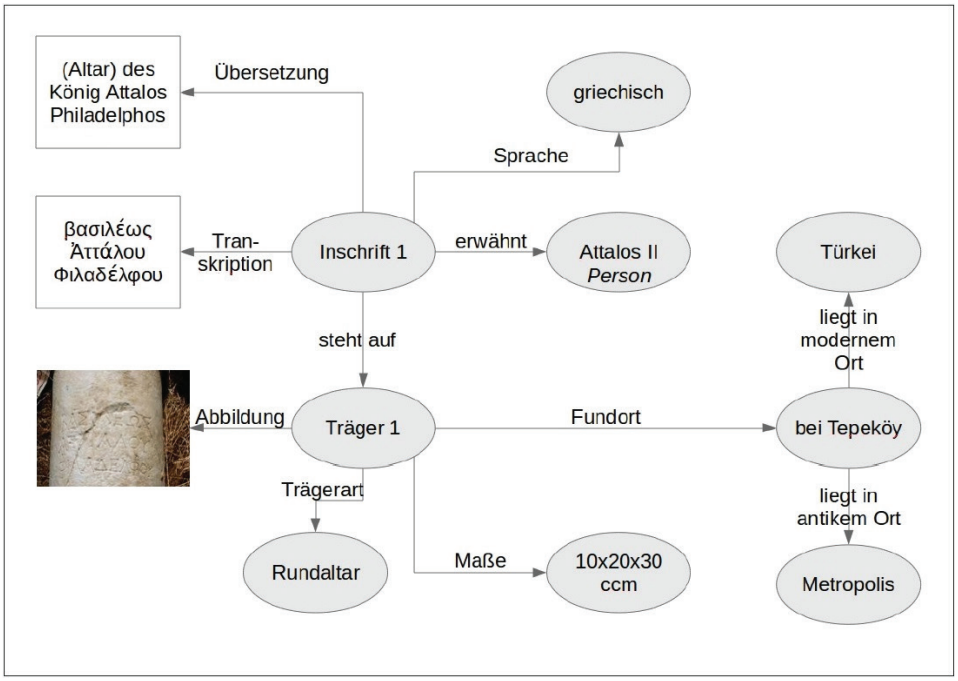

Abbildung 2: EDEN: Inschrift 1 Metropolis, Rundaltar des Königs Attalos II. Philadelphos, Darstellung als Netz

5 Anzeige und Eingabe

WissKI übernimmt das Prinzip von Wikipedia, dass jeder Artikel, also jede Seite, genau ein Objekt beziehungsweise ein Thema beschreibt. Das bedeutet, dass es in WissKI für jedes in den Daten beschriebene Objekt eine dezidierte Webseite gibt, die alle relevanten Informationen zum Objekt enthält. Diese Seite kann traditionell als „Datensatz" zum Objekt verstanden werden (vgl. Abbildung 1).

Da aufgrund der Netzstruktur auf der Datenebene kein solcher Datensatz besteht, bedarf es einer Entscheidung, welche Bereiche des Netzes als zum ›Datensatz des Objekts‹ gehörig betrachtet werden. Dies wird in WissKI durch die Definition von Eingabe- und Ansichtsmasken für die Objektart umgesetzt. EDEN erachtet beispielsweise für eine Inschrift die komplette Information zum Träger als relevant, während bei einem Träger lediglich Titel und Inhalt (Originaltext) der darauf vorhandenen Inschrift(en) angezeigt werden. Die Masken definieren also eine Abbildung der Netzstruktur in eine Datensatzstruktur und umgekehrt.

Durch die Kombination von datensatzorientierter Visualisierung und netzbasierter Wissensrepräsentation vereint die Datenbank die Vorteile beider Ansätze. Somit behält die Datenbank ihre Fähigkeit zur Verarbeitung komplexer Zusammenhänge und ermöglicht gleichzeitig einen intuitiven und unkomplizierten Zugang auf der Benutzerseite.

Auch der grundlegende Seitenaufbau orientiert sich an Wikipedia: Wie in Abbildung 1 zu sehen ist, steht im Zentrum Fließtext. Seitlich des Textes stehen Boxen mit den strukturierten Daten in tabellarischer Form sowie eventuelle Bilder. Über Links im Text und in den strukturierten Daten wird auf andere Objekte verwiesen. Zusätzlich können Informationen aus fremden Quellen wie dem ›Getty TGN‹[7] angezeigt werden, wenn denn die lokalen Daten damit verknüpft wurden. In EDEN werden beispielsweise die vorhandenen Geo-Koordinaten genutzt, um für Orte Thumbnails von ›Google Maps‹[8] zu generieren und zu verlinken. Im Rahmen der Forschungen zu Metropolis sind 2013 und 2014 zahlreiche 3D-Scans der Ausgrabungsstätten erstellt worden. Daraufhin wurden testweise 3D-Modelle als zusätzliche Darstellungs- und Zugangsmethode in EDEN (und erstmals in WissKI) eingeführt. Langfristiges Ziel ist einerseits die semantische Annotation der 3D-Objekte, ähnlich wie bei den Texten, und zum anderen die Verbindung der einzelnen Modelle in einer einzigen virtuellen Rekonstruktion der antiken Stadt.

Fließtext kann in mehrere Abschnitte untergliedert sein, die sich aus unterschiedlichen Einzeltexten zusammensetzen und die jeweils als eigene Textinstanzen innerhalb der Datenbank existieren. Der Fließtext aus Abbildung 1 setzt sich aus dem Originaltext der Inschrift – also dem eigentlichen Eintrag – sowie den textuellen Metadaten ›Deut-

7 Zu erreichen unter: `http://www.getty.edu/research/tools/vocabularies/tgn/`.
8 Zu erreichen unter: `https://www.google.de/maps/preview`.

sche Übersetzung‹, ›Englische Übersetzung‹, ›Regest‹ und ›Kommentar‹ zusammen. Wie bei den strukturierten Daten ist es also – anders als in Wikipedia – nicht nötig, Teiltexte mehrfach in verwandten Artikeln abzufassen beziehungsweise im ›copy-paste‹-Verfahren zu übertragen und später überall zu redigieren. Der entsprechende Abschnitt wird einfach den betreffenden Objekten zugeordnet und dort mit angezeigt.

WissKI bietet die Funktion, Annotationen im Fließtext vorzunehmen, aus denen dann strukturierte Daten automatisch extrahiert werden können. Somit entfällt eine zweifache Eingabe der Information sowohl im Text als auch bei den strukturierten Daten. Momentan ist die Funktion in EDEN auf Orte, Personen, Publikationen und Datierungen sowie andere Inschriften, die im Fließtext genannt werden, anwendbar. In Abbildung 3 (zu Illustrationszwecken vereinfacht) ist zu sehen, wie Daten aus einer Annotation im Fließtext automatisch extrahiert werden. In diesem Beispiel wird der Name ›Apollonis‹ durch Annotation im Text mit der entsprechenden Person verknüpft und dadurch automatisch die Relation zwischen Inschrift und Person dem Datenbestand hinzugefügt.

Abbildung 3: EDEN: Inschrift 2 Metropolis, Königin Apollonis von Pergamon, URL: `http://wisski.cs.fau.de/eden/content/ecrm_E34_Inscription01f3`

6 Suchen und Finden

Durch die starke Vernetztheit der Daten, die sich auf einer Seite visuell durch zahlreiche Verlinkungen ausdrückt, kann der Benutzer leicht den Datenbestand explorieren, ohne eine Suchfunktion in Anspruch nehmen zu müssen. Dabei spielt die oben genannte ›Randinformation‹ eine bedeutende Rolle: So ermöglicht EDEN den Zugang zu den Inschriften unter anderem über Texte, Gattungen, Orte, Datierungen und Personen.

Mittels Volltextsuche sind die griechischen und lateinischen Originaltexte sowie die wissenschaftlichen Kommentare zugänglich. Die Inschriften können auch nach anderen Kriterien systematisch durchsucht werden. So sollen beispielsweise alle Inschriften, die in

ein bestimmtes Jahrhundert datiert werden, ebenso angezeigt werden können wie diejenigen Inschriften, die noch nicht datiert sind. Ebenso können Personennamen und Fundstellen durchsucht werden. Durch Transkriptionen und Übersetzungen der Inschriften in verschiedene Sprachen sowie ausführliche Kommentare und erläuternde Hintergrundinformationen richtet sich EDEN neben den bereits angesprochenen Forschergruppen – Althistoriker und Archäologen – auch an ein breiteres Publikum: Es ist für Studierende und interessierte Laien ein wertvolles Nachschlagewerk und für die Lehre einsetzbar.

7 Planen einer semantischen Datenbank mit WissKI

Die vorhergehenden Abschnitte haben einige Vorzüge des WissKI-Systems beleuchtet. Dieser Ansatz verlangt aber auch einiges Umdenken im Vergleich zu vielen anderen Software-Werkzeugkästen. Zum einen erfordert ein reines Online-System, sich Gedanken über Serverplatz und -wartung zu machen, um eine ständige (weltweite und dauerhafte!) Verfügbarkeit zu gewährleisten und damit die Arbeitsgrundlage sicherzustellen. Zudem sind Sicherheitsaspekte und Workflows zu beachten bzw. zu definieren, die auf einem lokalen Arbeitsplatzrechner durch die physische Begrenztheit vorgegeben sind. Diese Aspekte sind allerdings nicht neu und mittlerweile schon Alltag bei Projektplanungen. Hier wird daher auf die semantische Anreicherung der Daten in WissKI eingegangen, die gleichzeitig die Beziehung festlegt zwischen der Datenrepräsentation in der Datenbank und ihrer Präsentation auf der Weboberfläche. So steht auch beim Einsatz von WissKI am Anfang die Frage: Wie bereite ich mich vor? Hier kommen einige Erfahrungen beim Aufbau von EDEN und anderen WissKI-Anwendungen zur Sprache.

Zunächst ist WissKI auch eine Datenbank. Beim Aufbau einer Datenbank wird man typischerweise gleich zu Beginn mit der Frage konfrontiert: Welche Felder brauche ich? Die Antwort auf diese Frage ist von mehreren Faktoren abhängig, etwa der Forschungsdisziplin, dem Erkenntnisinteresse, dem Detailgrad der Objekt- und Textbeschreibungen sowie der Anwendungsabsicht. Gerade letzterer Punkt wird häufig zu wenig explizit betrachtet, beeinflusst aber stark die Datenerfassung und -darstellung und damit die Wahl der Felder. Daher sollen hier noch zwei wichtige Fragen genannt werden: Welche Zusammenhänge will ich darstellen? und Welche Fragestellungen will ich unterstützen?

Generell gilt, je effektiver das Grundgerüst vorab geplant wurde, desto weniger Nachbesserungsarbeit fällt später an. Dennoch wird man beim Erstellen einer semantischen Datenbank mit großer Wahrscheinlichkeit hin und wieder auf Fälle stoßen, die eine weitere Ausdifferenzierung der Felder erfordern. In EDEN wurden beispielsweise die Felder zu den Personendaten erst später hinzugefügt und auch die Differenzierung zwischen antiken und modernen (Fund-)Orten ist ein späterer Zusatz. Mit WissKI kann dies allerdings problemlos parallel zum Einpflegen der Daten erfolgen. Somit kann die

Datenbank nicht nur hinsichtlich der Datenmenge, sondern auch hinsichtlich der Fülle der Metadaten bei Bedarf stetig wachsen.

In WissKI ist der Aufbau der Eingabemasken und -felder visuell bewusst an Karteikarten beziehungsweise Eingabemasken für relationale Datenbanken angelehnt. Die Planung und Erstellung solcher Masken ist also in dieser Hinsicht Routine. Der entscheidende Unterschied bei WissKI ist die Abbildung der Bedeutung eines Datenfeldes auf die zugrundeliegende Ontologie. Dies erfordert zum einen ein klares Verständnis der Ontologie und ihrer Konstrukte. WissKI empfiehlt hier das ›CIDOC CRM‹[9]: Es bietet ein ausgereiftes, generisches sowie erweiterbares Modell, und durch die vordefinierten, generischen Kategorien braucht das Rad nicht jedes Mal neu erfunden zu werden. Zum anderen bedingt dies, dass die Bedeutung des Feldes klar umrissen und definiert sein muss, um überhaupt auf die Ontologie abbildbar zu sein. Auf den ersten Blick mag das wenig aufwendig erscheinen; jedoch ist gerade bei der Verwendung von bereits existierenden Datenbanken die Frage nach der genauen Semantik der Felder nicht immer leicht zu beantworten: So machen fehlende Dokumentation und mehrere Generationen von Dateneingebern mit unterschiedlichen Vorstellungen von der Bedeutung des Feldes es manchmal sehr schwierig, diese klar zu bestimmen.

Dies mag sich nun aufwendig und abschreckend anhören. Doch der augenscheinlich höhere Aufwand der Erschließung der Semantik der Daten birgt auf mittlere oder lange Sicht klaren Mehrwert: Nicht nur werden die Daten durch die technischen Möglichkeiten des Semantic Web leicht vernetzbar gemacht, sowohl lokal als auch weltweit. Das Vergegenwärtigen und Erfassen der Bedeutung führt auch zu einer vertieften Reflexion der eigenen Tätigkeit und des eigenen Datenbestandes und kann letztendlich helfen, die Qualität der Daten zu erhöhen. Diese Erfahrung macht einer der Autoren immer wieder bei Diskussionen mit WissKI-Anwendern, so auch bei EDEN.

9 ISO 21127; siehe auch: http://www.cidoc-crm.org/.

ANNOTATIONEN OHNE ENDE?
AUSZEICHNUNGSPROZESSE AM BEISPIEL DER REGESTA PONTIFICUM ROMANORUM ONLINE

Klaus Herbers, Thorsten Schlauwitz

Zunehmend wird die Forderung nach einer digitalen Publikation von Forschungsergebnissen an Wissenschaftsprojekte herangetragen[1]. Die Vorteile sind evident: Die Daten stehen weltweit und meist kostenfrei zur Verfügung. Zudem ist eine nachträgliche Ergänzung bzw. Korrektur möglich, was bei der Druckfassung nur mittels einer kostspieligen Neuauflage verwirklicht werden kann. Nicht zuletzt verbessern aber digitale Publikationen die Recherchemöglichkeiten. Das trifft bereits auf reine pdf-Publikationen zu, in denen eine elektronische Suche schneller ist als die Arbeit mit einem Register. Umso mehr gilt dies für Datenbanken, deren Suchparameter meist nicht nur die Suchmethoden eines Registers verfeinern und erweitern, sondern durch kombinierbare Abfragen auch neue Zugriffsweisen bieten. Diese komfortablen Angebote für den Nutzer sind aber nicht ohne teils erheblichen Arbeitsaufwand seitens der Bearbeiter zu erreichen. Es ist daher ein stetes Postulat, die notwendigen Arbeitsschritte zu vereinfachen und zu automatisieren, um derartige Funktionen zur Verfügung stellen zu können. Eine Optimierung dieser Prozesse trägt dazu bei, dass auch bei online-Publikationen ein einheitlicher Standard gewährleistet bleibt. Es existieren Beispiele, bei denen dies nicht der Fall ist, wodurch die betroffenen Datenbanken bzw. die spezifischen Suchparameter ebenso an Wert verlieren wie ein Register, welches 50 Seiten eines Buches nicht erschlossen hat. Erschwerend kommt hinzu, dass auf diese Ungleichgewichte häufig nicht hingewiesen wird. Im Folgenden werden die Lösungsmöglichkeiten, die für eine Datenbank mediävistischer Quellen (*Regesta Pontificum Romanorum online [RPR]*, **www.papsturkunden.de**) entwickelt wurden und für zukünftige Projekte eine Hilfe darstellen können, vorgestellt. Dabei wird auch erläutert, wie man die Vorteile einer modernen XML-Datenbank und einer relationalen Access-Datenbank miteinander kombinieren kann.

1 Vgl. dazu die Berliner Erklärung zu Open Access, **http://openaccess.mpg.de/Berliner-Erklaerung**, letzter Zugriff am 6.11.2014.

Papsturkunden bis zum Jahr 1198

Das hier behandelte Vorhaben widmet sich seit 1896 der Verzeichnung der Papsturkunden bis zum Jahr 1198. Deren Erforschung gehört zu einer der zentralen Aufgaben der Mediävistik. Das Papsttum als eine der beiden universalen Anspruch erhebenden Mächte des Mittelalters neben dem Kaisertum hatte Kontakte zum gesamten Abendland und nahm mit seinen Entscheidungen auch Einfluss auf zahlreiche Bereiche, die heute der kirchlichen Sphäre entzogen sind. Seine Bedeutung manifestiert sich besonders in den Urkunden der Nachfolger Petri: sowohl bezüglich der Quantität als auch der Qualität war die päpstliche Kanzlei die leistungsfähigste ihrer Art im Mittelalter.

Die Zusammenstellung der ausgehenden Briefe und Urkunden gestaltet sich als äußerst zeit- und arbeitsaufwändig, da sich die Urkunden vor dem Scheidejahr 1198, dem Beginn der kontinuierlichen Registerführung an der Kurie, nur bei den Empfängern, nicht aber in Rom selbst erhalten haben. Daher sind für diese frühere Zeit umfassende Recherchen in den Archiven Europas notwendig. Neben dem Sammeln ist auch die Präsentation der Urkunden eine wichtige Aufgabe. Um einen schnellen Überblick über die Schreiben zu erhalten – mittlerweile geht man von insgesamt über 30.000 Papstkontakten vor 1198 aus – wird der rechtsrelevante Inhalt jeder Urkunde kurz zusammengefasst. Diese sogenannten Regesten sind außerdem mit Angaben zu Überlieferung, Edition und einem Sachkommentar versehen.

Drei Projekte – eine Datenbank

Der Bedeutung dieser Quellengruppe entsprechend, haben sich drei große Projekte der Erschließung der Papsturkunden gewidmet. Die älteste Zusammenstellung stammt von Philipp Jaffé, bei der es sich um ein rein chronologisch sortiertes Verzeichnis bis 1198 handelt, welches 1885–87 eine zweite Auflage erfuhr. Daneben erarbeitet das 1896 von Paul Fridolin Kehr ins Leben gerufene Göttinger Papsturkundenwerk Regestenbände nach einem geographisch-institutionellen Ordnungsprinzip. Schließlich widmen sich auch die Regesta Imperii, die sich zunächst auf die Königsurkunden konzentrierten, dann aber die Bedeutung der Papsturkunden für die Reichsgeschichte erkannten, verstärkt seit ca. 1950 dieser Aufgabe. Während die letzten beiden Reihen trotz erheblicher Fortschritte bisher noch nicht abgeschlossen sind, wird die über hundert Jahre alte Fassung der zweiten Auflage des Jaffé derzeit in Erlangen neu bearbeitet. Die unterschiedlichen Bearbeitungsmaßstäbe und der von Band zu Band variierende Bearbeitungszeitpunkt zwingen die heutigen Mediävisten, regelmäßig alle drei Reihen zu konsultieren. Die teils fehlenden Register erschweren dabei die Erschließung dieser Bände.

Diese Schwierigkeiten werden durch die im November 2013 zur Verfügung gestellte Online-Datenbank *Regesta Pontificum Romanorum online* behoben, die im Rahmen des Göt-

tinger Akademienprojektes ›Papsturkunden des frühen und hohen Mittelalters‹ (http://www.papsturkunden.gwdg.de) in Erlangen entwickelt wurde. In dieser wird zukünftig für jeden belegbaren Papstkontakt vor 1198 ein Regest vorhanden sein. Es können zudem alle drei genannten Projekte in die Datenbank integriert werden, so dass zu einem Papstkontakt das jeweilige Regest aus jeder der drei Reihen parallel konsultierbar ist. Allein durch die Aufhebung dieser institutionell bedingten Aufgliederung wird den Forschern eine erhebliche Arbeitserleichterung an die Hand gegeben.

Funktionen der Datenbank

Das Angebot wird gegenüber den jeweiligen Druckfassungen erheblich erweitert. Durch Korrekturen und Ergänzungen können die Daten stets aktuell gehalten werden. Während zunächst allein die neu erscheinenden Bände zeitnah in die Datenbankstruktur eingebunden werden, werden die älteren Bände des Papsturkundenwerkes schrittweise als PDF-Dokument (OCR-basiert) zur Verfügung gestellt. Durch eine seitengenaue Verlinkung zwischen dem einzelnen Datensatz und den PDF-Dokumenten können einerseits die älteren Regesten leicht konsultiert werden, andererseits auch die aktuelle Fassung des Regests mit der ursprünglichen Druckfassung verglichen werden, wodurch ein steter Abgleich zwischen der derzeitigen Datenbankfassung und der ursprünglich gedruckten Version erreicht wird. Zudem können diese Bände auch als Ganzes gelesen werden, was u. a. bezüglich der historischen Einleitungen zu den Institutionen aus den Pontificia-Bänden hilfreich ist.

Auf diese Weise können aber nicht nur die Regesten der drei Projekte, sondern auch beispielsweise Editionen (zunächst die des Papsturkundenwerkes) verlinkt werden, wodurch die inhaltliche Erschließung in Form des Regests sowie der Volltext zusammengeführt werden. Daneben können weiterhin Abbildungen der Papsturkunden den einzelnen Datensätzen angefügt werden, wobei hier vor allem auf die umfangreiche Sammlung der Göttinger Arbeitsstelle zurückgegriffen wird, aber auch Verlinkungen auf andere Angebote wie beispielsweise monasterium.net und das Marburger Lichtbildarchiv die Möglichkeiten erweitern. Durch dieses breite Angebot wird nicht nur die Bearbeitung historischer, sondern auch diplomatischer, paläographischer und linguistischer Fragestellungen möglich.

Neben der Verknüpfung der verschiedenen Informationen stellen aber vor allem die technischen Möglichkeiten der modernen XML-Datenbank (basierend auf einer eXist-Datenbank, die Eingabe erfolgt über ein Java-Applet) einen Zugewinn dar. So können die Daten durch wesentlich vertiefte Suchparameter erschlossen werden, die neben einer Volltextrecherche eine spezifische Suche nach fast 30 verschiedenen Kriterien erlauben. Während ein Großteil dieser Suchmöglichkeiten durch eine separierte Speicherung in

verschiedenen Feldern – oder besser gesagt, in eigenen XML-Tags – ermöglicht wird, ist es besonders wichtig, die zwei Informationen ›Personen‹ und ›Orte‹ zu erschließen. Dies ist aber unabhängig von den inhaltlichen Identifizierungsproblemen auch technisch schwer umsetzbar. Da ist zunächst die Problematik, dass diese Informationen sich in mehreren Feldern (Regest, Sachkommentar, Unterschriften usw.) befinden. Daneben existieren sprachliche Barrieren. Während die Regesten der Regesta Imperii grundsätzlich auf Deutsch mit lateinischen Quellenzitaten verfasst werden, sind die beiden anderen Regestenwerke ausschließlich in lateinischer Sprache. Weiterhin werden Eigennamen in den Regesten nicht normalisiert, sondern nach der Schreibweise in den jeweiligen Quellen aufgenommen. Dieses Vorgehen erschwert es aber dem Datenbankbenutzer, alle Treffer zu einer Person oder einem Ort zu finden, da er alle möglichen Schreibweisen prüfen müsste. Hinzu kommt, dass bei den Personen eine Verfeinerung nach der Funktion als Aussteller, Adressat oder Empfänger der Urkunde vorgenommen wird.

Zur Bewältigung dieses Problems müssen alle Personen- und Ortsnamen mit normierten Daten hinterlegt werden. Bei Personen werden neben dem normalisierten Namen (Name in der jeweiligen modernen Landessprache ohne Sonderzeichen) das Todesdatum, der Wirkungsort (bspw. Bischofssitz), Namenszusätze (»der Große«) sowie eine eindeutige ID, die von einer Personendatenbank zur Verfügung gestellt wird und über eine Verlinkung weitere Informationen zu den jeweiligen Personen bereit hält (dem internationalen Rahmen des Projektes entsprechend, wurde hier die Virtual International Authority File, **www.viaf.org**, der ansonsten in Deutschland primär genutzten GND vorgezogen), aufgenommen. Zukünftig können dadurch über das Beacon-Format auch Treffer zu der Person in anderen Datenbanken angezeigt werden. Bei den Orten wird neben dem normalisierten Namen eine Kategorisierung der Institution (Stadt, Bistum, Kloster) durchgeführt sowie der Name der Institution, die Diözese und ebenfalls eine eindeutige ID, in diesem Fall die von **www.geonames.org**, festgehalten.

Diese Tags wurden zunächst über eine Benutzeroberfläche mittels Auszeichnungen in der XML-Datenbank vorgenommen, die Normdaten konnten als Attribut eingetragen werden (vgl. Abb. 1). Ein ähnliches Verfahren, wenn auch mit einer anderen Zielsetzung, wird bei den Literaturtiteln angewendet: Um nicht jedes Mal vollständige Literaturtitel zitieren und um langfristig nicht mit der Einheitlichkeit von Zitierstilen kämpfen zu müssen, werden in der RPR nur Kurztitel verwendet, die auf den in der Mediävistik verbreiteten OPAC der Regesta Imperii verweisen. Deshalb muss zu jedem Literaturtitel der entsprechende Link hinterlegt werden.

Dieses Verfahren wurde bei den ersten Datensätzen, den 287 Regesten der Bohemia Pontificia[2], komplett in der XML-Datenbank umgesetzt. Da hierfür jede Person, jeder Ort

2 Waldemar Könighaus: Bohemia-Moravia Pontificia vel etiam Germania Pontificia V/3: Provincia Maguntinensis. Pars VII: Dioeceses Pragensis et Olomucensis, Göttingen 2011.

und jeder Literaturtitel einzeln markiert und die Normdaten manuell eingefügt werden mussten, hat sich dies schnell als ein äußerst arbeits- und zeitaufwändiges Verfahren erwiesen, welches zu optimieren war. Zudem birgt dieses Verfahren die große Gefahr von Tippfehlern, da hier die Informationen jeweils separat eingegeben werden und nicht wie in einer relationalen Datenbank auf die Daten in einer hinterlegten Tabelle zurückgegriffen wird. Deshalb wurde der Arbeitsprozess umgestellt und verbessert.

Arbeitsumgebung/Import

Zur Erleichterung dieses Arbeitsschrittes trug der bereits zuvor etablierte Schritt des Imports bei. Die Regesten wurden nie direkt online erstellt, sondern seit Beginn des Göttinger Akademienprojektes in lokalen MS Access-Datenbanken bearbeitet. Dies ist aus mehreren Gründen vorteilhaft. MS Access bietet einerseits eine (relativ) einfache Benutzeroberfläche, welche die Dateneingabe und -pflege, aber auch die Programmierung ohne (vertiefte) Kenntnisse einer Programmiersprache erlaubt. Vorhanden sind aber auch alle technischen Optionen (z. B. Filterung, kombinierbare Suchoptionen), die für die meisten geisteswissenschaftlichen Projekte zentral sind. Zudem ermöglicht diese Vorgehensweise, die lokalen Datenbanken den individuellen Bedürfnissen der jeweiligen Bearbeiter durch beispielsweise überarbeitete Eingabemasken anzupassen. Weiterhin können lokale Datenbanken problemlos weltweit, auch bei Archivreisen, verwendet werden, während man sonst auf einen Internetzugang angewiesen wäre. Auch der Wechsel zwischen den Datensätzen, die Sortierung und das Filtern fallen in der Access-Datenbank leichter als dies bei den noch nicht freigegebenen Regesten in der XML-Datenbank der Fall ist. Zuletzt kann aus den lokalen Datenbanken mittels eines Seriendruckes verhältnismäßig einfach und schnell ein Word-Dokument zur Vorbereitung der Drucklegung generiert werden. Selbst eine Zusammenarbeit mit dem immer weitere Verbreitung findenden Textsatzprogramm LaTex ist möglich. Zur Transferierung der Daten aus den Access-Datenbanken in RPR wurde ein automatisierendes Script geschrieben, wodurch sich mehrere hundert Datensätze problemlos in wenigen Minuten in die Datenbank integrieren lassen. In diese Import-Funktion ist zudem eine Dublettenprüfung integriert, die gegebenenfalls verschiedene Regesten zu einem Papstkontakt automatisch zusammenführt beziehungsweise den Import von bereits vorhandenen Regesten ablehnt.

Zwischen der Bearbeiterdatenbank und der XML-Datenbank musste allerdings eine dritte Datenbank, eine lokale Access-Datenbank, in den Importprozess eingebunden werden. In dieser ›Importdatenbank‹ (Import-DB) werden die leichten Unterschiede zwischen den verschiedenen Benutzerdatenbanken von mittlerweile über einem Dutzend Bearbeitern ausgeglichen, um damit die Daten für den Import in die RPR zu vereinheit-

lichen. Durch die Generierung verschiedener Abfragen kann dies automatisiert werden, indem beispielsweise Feldinhalte einer Spalte getrennt bzw. zusammengefügt werden.

Auszeichnungen

Innerhalb dieser Import-DB werden auch die Auszeichnungen vorgenommen, da sowohl die entsprechenden Anwendungen in der XML-Datenbank (vgl. Abb. 1) als auch die manuelle Eintragung der XML-Tags ein zu großer Aufwand wäre (vgl. das Beispiel in Abb. 2).

Stattdessen greifen an dieser Stelle die Vorteile einer relationalen Datenbank. Der Grundgedanke ist dabei, dass die Normdaten nur einmalig angelegt werden und anschließend durch eine Suchen-Ersetzen-Prozedur der jeweilige Begriff durch den entsprechenden XML-Code ausgetauscht wird. Dieses Verfahren weist zwei Vorteile auf: Die Normdaten zu einer Entität sind identisch und die Gefahr von Tippfehlern wird erheblich reduziert. Weiterhin können die eingetragenen IDs zu den externen Datenbanken mittels eines eingebundenen Webbrowsersteuerelements, welches die entsprechende Internetseite im Eingabeformular anzeigt, direkt überprüft werden. Außerdem wird der Zeitaufwand deutlich reduziert. Prinzipiell sind drei verschiedene Tabellen notwendig. In einer ersten Tabelle werden die Normdaten (Attribute) eingefügt. In einer zweiten, damit verknüpften Tabelle müssen die ›Quellbegriffe‹ verzeichnet werden, also sämtliche Schreibvariationen, wie sie in den Regesten auftreten können. Diese beiden Tabellen stehen in einer 1:n-Beziehung. Zu jedem Normnamen können somit mehrere ›Quellbegriffe‹ eingetragen werden, womit den verschiedenen Schreibweisen und Flexionsformen Rechnung getragen wird. In einer dritten Tabelle werden schließlich die eigentlichen Papstregesten gespeichert. Durch eine Aktualisierungsabfrage werden die Suchbegriffe durch den entsprechenden XML-Tag ersetzt. Um Falschauszeichnungen zu vermeiden, wird beim Suchprozess nicht ausschließlich nach dem Suchbegriff gesucht, sondern es werden automatisch Leerzeichen bzw. Satzzeichen am Anfang und Ende ergänzt (statt nach dem eingetragenen »Roma« wird nach » Roma «, » Roma.«, » Roma, «, » Roma: « und » Roma; « recherchiert; beim Ersetzen werden diese Satzzeichen ebenfalls ergänzt).

Auch dieses Verfahren hat Schwächen. Es ist rechnerintensiv, so dass ›nur‹ wenige hundert Datensätze auf einmal bearbeitet werden können. Grundsätzlich ist zudem wie bei allen Automatisierungsprozessen ein manueller Korrekturgang notwendig, um gegebenenfalls falsche Auszeichnungen wieder zu entfernen. Während das Verfahren für die Ortsnamen und Literaturdaten dennoch weitgehend einwandfrei funktioniert, bleiben bei den Personennamen noch einige Probleme. Ursache hierfür sind die Quelltexte, die Regesten. Die dort genannten Personen haben im Gegensatz zu den Literaturtiteln und den Ortsnamen häufig keinen einzigartigen ›Quellbegriff‹, nach dem gesucht werden kann. Erinnert sei nur an den häufigen Personennamen *Johannes*. In diesen Fällen ist

daher eine manuelle Vor- oder Nachbearbeitung notwendig. Dennoch ist auch hier im Vergleich zu einem Auszeichnungsprozess innerhalb der XML-Datenbank der Arbeitsaufwand geringer.

Dieses mehrstufige, auf die individuellen Anforderungen des Papsturkundenprojektes zugeschnittene Verfahren kann auch für andere Vorhaben adaptiert werden. Insgesamt erweist sich besonders bei Projekten mit längerer Laufzeit die stete Suche nach Automatisierungsprozessen als ertragreich. Hierfür genügen besonders für projektinterne Arbeitsschritte häufig weit verbreitete Standard-Softwarepakete, deren Anwendung in vergleichsweise geringer Zeit erlernt werden kann.

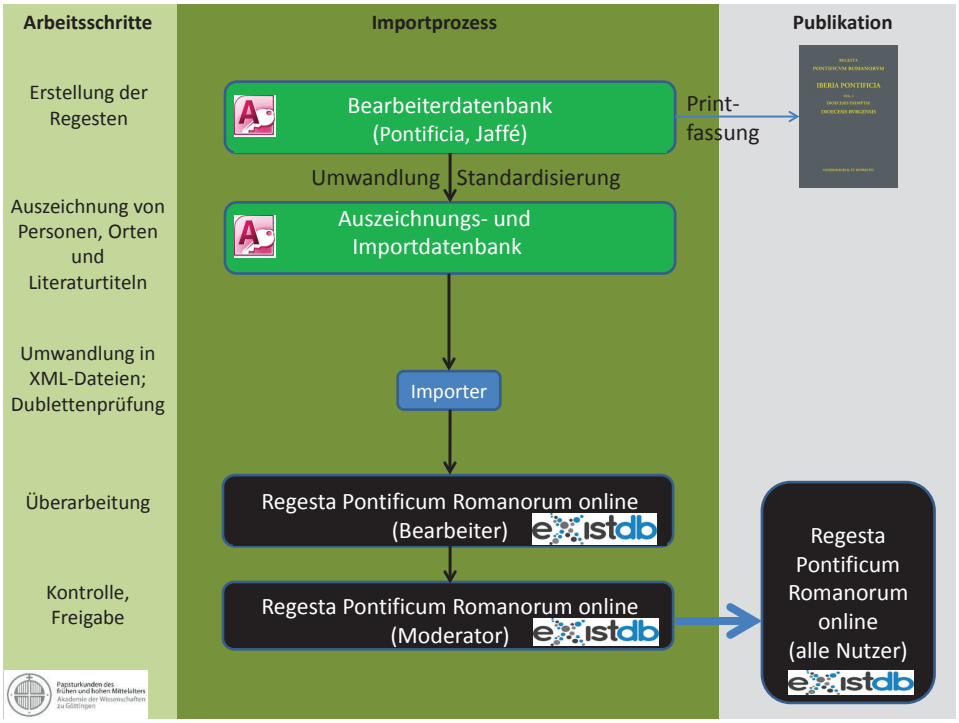

Abbildung 1: Bearbeitungsoberfläche der XML-Datenbank

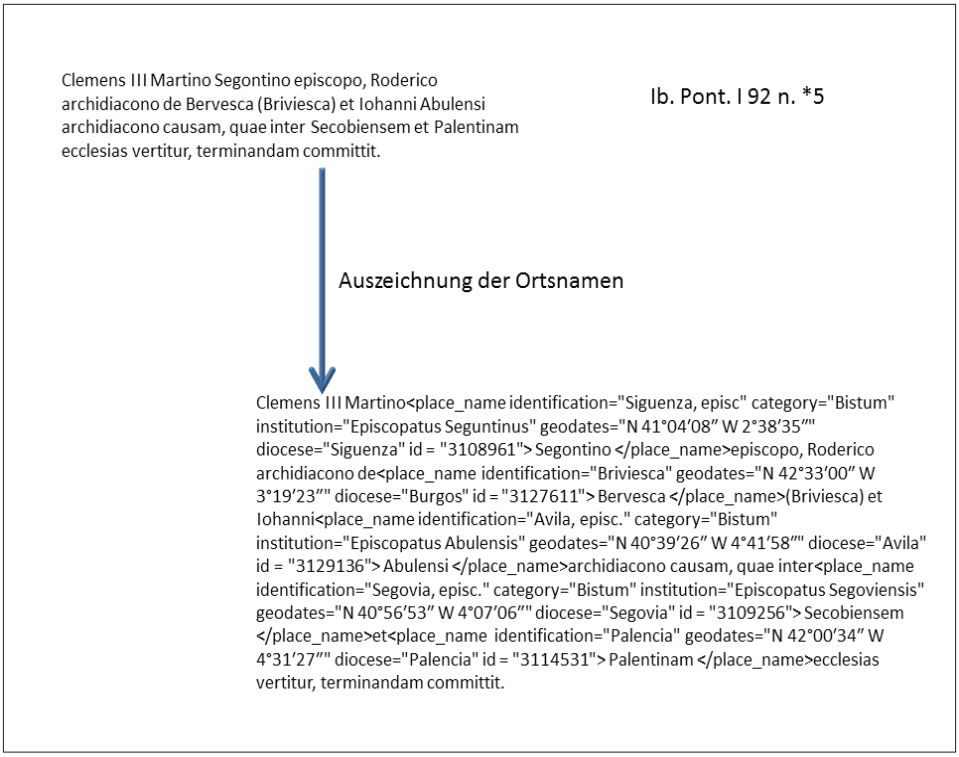

Abbildung 2: Auszeichnung eines Regests mit Ortsnamen

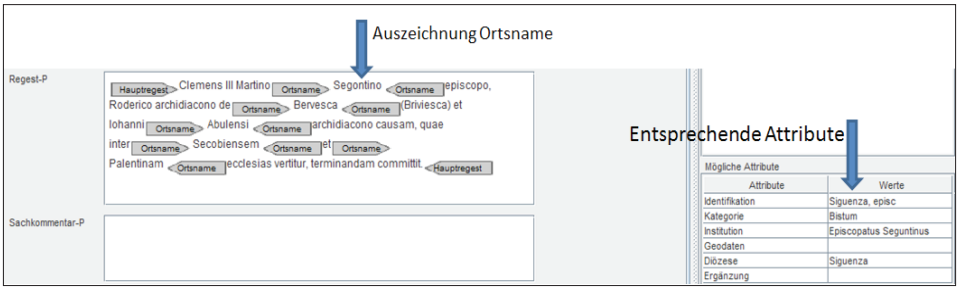

Abbildung 3: Importprozess