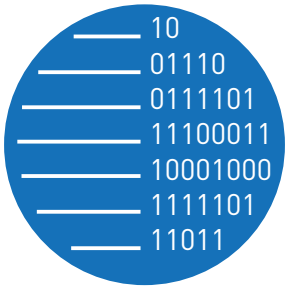

MAGAZIN FÜR DIGITALE EDITIONSWISSENSCHAFTEN

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

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.

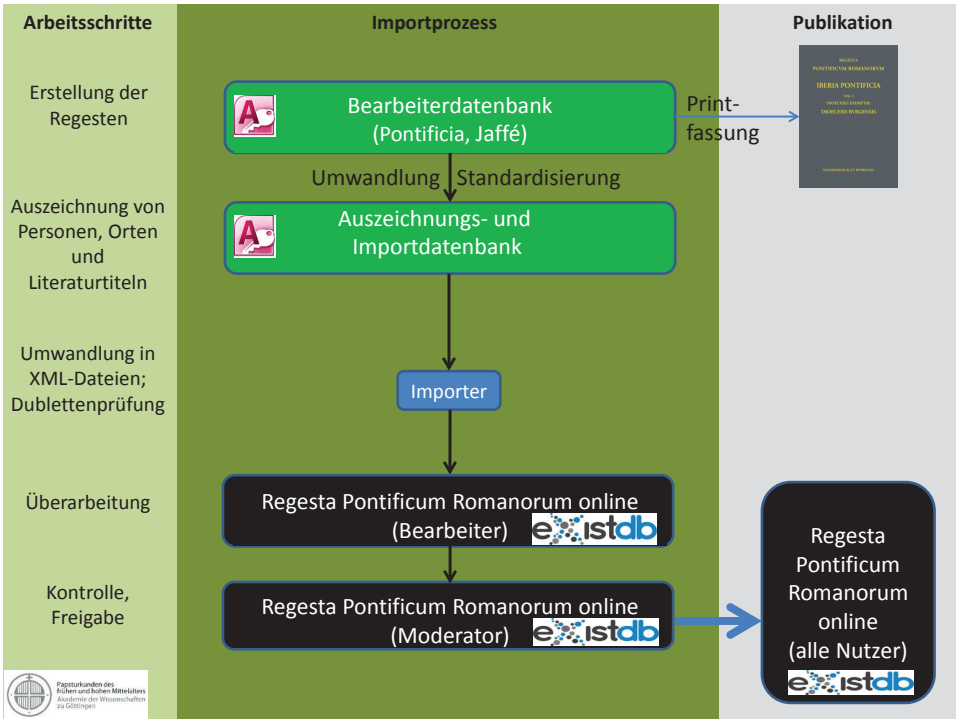

Abbildung 1: Bearbeitungsoberfläche der XML-Datenbank

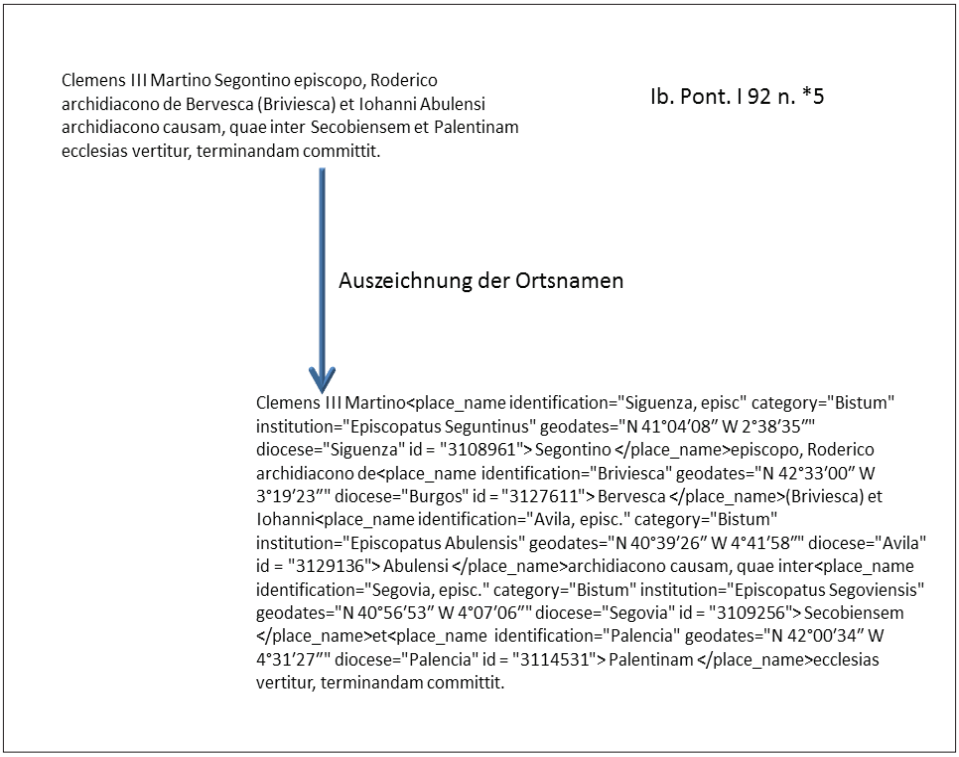

Abbildung 2: Auszeichnung eines Regests mit Ortsnamen

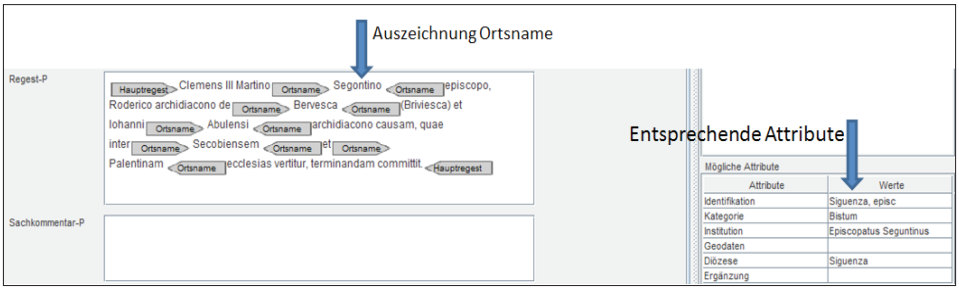

Abbildung 3: Importprozess