Big Data History

Andreas Kuczera

10.10.2014

1 Der Blick zurück – als ich promovierte

Als ich Ende der 90er Jahre des letzten Jahrtausends meine Promotion zur Grundherrschaft eines Klosters in Mittelhessen begann, stand am Anfang die systematische Untersuchung der Überlieferung. Alle für den untersuchten Zeitraum relevanten Urkunden habe ich mindestens einmal gelesen und den Inhalt auf Relevanz für meine Arbeit untersucht. Zu dieser Zeit waren digitalisierte Volltexte und Quellen noch die Ausnahme. Heute, gut 15 Jahre später, stehen Historiker vor einer anderen Situation. Sie haben Zugriff auf eine immer weiter wachsende Anzahl an Volltexten und digitalen Quellen, Metadaten usw.

Bei der Digitalisierung lag der Fokus zu Beginn noch auf der Imagedigitalisierung mit entsprechenden Metadaten. Selbst die MGH stellten ihre digitalisierten Buchseiten zunächst als Scans im Netz bereit, ohne direkten Zugriff auf die Volltexte zu bieten. Auch die Regesta Imperii, die von 2001 bis 2006 im Rahmen eines DFG-Projekts digitalisiert wurden, standen vor der Frage, wie die Texte im Netz dargeboten werden sollten[1]. Ich selbst war damals Mitarbeiter in diesem Projekt

[1] Zur Entwicklung des frühen Digitalisierungsprojekts vgl. Kuczera, Andreas: Die Regesta Imperii Online (2007) - In: Historisches Forum Bd. 10 (2007). Zum aktuellen Stand vgl. Weller, Tobias: Die Regesta Imperii Online (2014) - In: Rheinische Vierteljahrsblätter Bd. 78 (2014) S. 234-241; Würz, Simone: Mittelalterliche Quellen im Internet: Aspekte der Digitalisierung und Vernetzung der Regesta Imperii Online (2011) - In: Archive im Web - Erfahrungen, Herausforderungen, Visionen S. 162-171

und wir entschieden uns für eine Volltextdarstellung, die dem Buch möglichst nahe kommen sollte. Vor allem aber war eine Volltextdarstellung über HTML wesentlich leichter zu implementieren als die Präsentation von Scans mit dahinter verstecktem Volltext, wie sie zeitweise von den MGH angeboten wurde.

2 Neue Rezeptionsmöglichkeiten und der Ausgleich buchtechnischer Nachteile

Der DFG-Antrag hob u.a. hervor, dass mit dem Digitalisierungsvorhaben zum einen buchtechnische Nachteile ausgeglichen und zum anderen neue Rezeptionsmöglichkeiten erschlossen werden sollten. Dabei bezog sich der Hinweis auf buchtechnische Nachteile beispielsweise auf die Abteilungen 7 (Ludwig der Bayer) und 13 (Friedrich III.), die im Unterschied zu den "regulären" Regesta Imperii-Bänden nicht die komplette Überlieferung, sondern jeweils mit einem Heft den Quellenbestand eines Archivs oder einer Archivlandschaft enthalten. Dies führt dazu, dass man für die Recherche eines bestimmten Zeitraumes parallel alle bisher publizierten Hefte durchsehen muss, was offensichtlich großen Arbeitsaufwand mit sich bringt. Heute sind solche Recherchen mit einer Suchanfrage im Regestenmodul wesentlich leichter möglich. Auch die Bereitstellung einer Volltextsuche über die Regesta Imperii hat sicherlich den einen oder anderen Historiker auf Spuren gebracht, die er mit der alleinigen Analyse der gedruckten Bände möglicherweise nicht entdeckt hätte.

3 Nutzerverhalten und Nutzerperspektive

Im Rahmen eines Vortrages auf der "Digital Diplomatics 2013" im November letzten Jahres in Paris stellten meine Kollege Torsten Schrade und ich u.a. einige Analysen zum Nutzerverhalten des Regestenmodules der Regesta Imperii Online vor.

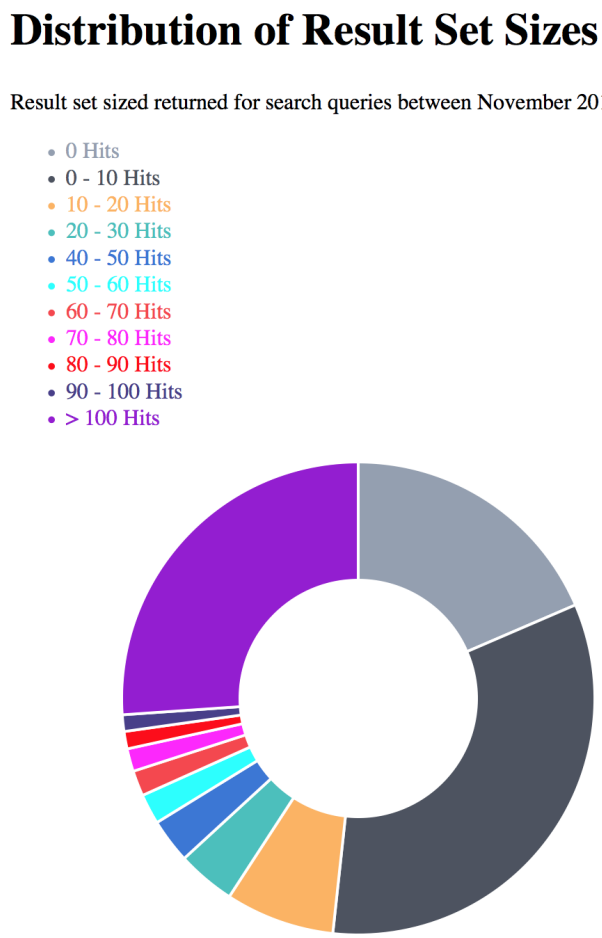

Abbildung 1: 10er-Schritte Anzahl der Treffer im Regestenmodul der Regesta Imperii Online (Quelle: Digitale Akademie, Mainz – www.digitale-akademie.de)

Datengrundlage waren die anonymisierten Daten von Suchanfragen aus dem Zeitraum von November 2012 bis Ende Oktober 2013. Bei jeder Suchanfrage wurde auch die Anzahl der erzielten Treffer mitgeloggt. Die Abbildung zeigt nun in einem Tortendiagramm wieviele der ca. 101.000 Nutzeranfragen im Regestenmodul 0 bis 10 Treffer, wieviele 11 bis 20, 21-30 usw. bis zu mehr als 100 Treffer erzielt haben. Es zeigt sich unter anderem, dass sich die Nutzer des Regestenmodules von ihren Ergebnissen her grob in drei Gruppen einteilen lassen. Die erste Gruppe nutzt die Expertensuche des Regestenmoduls optimal aus und bekommt in der Regel auf eine Anfrage zwischen 1 und 10 Regesten zurückgemeldet. Diese können dann am Bildschirm gelesen, gedruckt oder sonst weiter verarbeitet werden. Eine zweite Nutzergruppe bekommt zwischen 10 und 100 Treffern. Die dritte Gruppe erhält auf ihre Anfrage 100 oder mehr Treffer.

Erhält man bei der Regestensuche 1 bis 10 Treffer ist das Ergebnis mit vertretbarem Aufwand lesbar und zu überprüfen. Bei über 10 bis 100 Treffern würde ich vermuten, dass diese Nutzer versucht sein könnten, ihre Suchkriterien zu verschärfen und damit ein besseres (in diesem Fall auch kleineres) Ergebnis zu erhalten. Die dritte Gruppe erhält 100 oder mehr Treffer, deren Auswertung am Bildschirm äußerst mühselig ist.

Zu der letzten Gruppe gehören oft auch Nutzer, die mit einen einzigen Suchbegriff ohne weitere einschränkende Angaben sehr viele Treffer zurückgeliefert bekommen. Diese Gruppe hat aus meiner Sicht den "Google-Anspruch" bei minialem Input in die Suchmaske optimale Ergebnisse zu erhalten. Selbstverständlich kann man hier einwenden, dass Nutzer eines Quellenportals zumindest rudimentäre inhaltliche Kenntnis des untersuchten Gegenstandes, hier also der Regesta Imperii, mitbringen sollten.

Was mich bei den Ergebnissen aber erstaunte, war die Nutzungsform der Regestensuche. Die meisten Nutzer wussten genau, was sie suchten. Sie wählten die Bandansicht, riefen einen Band auf und suchten sich das gewünschte Regest. Sie nutzten die Onlineregesten in der gleichen Weise wie einen gedruckten Regestenband – nur kamen sie schneller ans Ziel. Den neuen Rezeptionsmöglichkeiten,

wie einst im DFG-Antrag formuliert, entsprach dies aber sicherlich nicht.

Bei der Diskussion im Kollegenkreis über die Ergebnisse war der Hinweis auf die mangelnde inhaltliche Kompetenz der Nutzer ein häufiger Reflex. Zunächst reagierte ich ebenso und machte die fehlende Kenntnis über die Regesta Imperii für die hohen Treffermengen verantwortlich. Dann aber fiel mir auf, dass die Nutzer mit "Google-Anspruch" mit ihren hohen Treffermengen vielleicht einfach eine neue Nutzungsform unseres Online-Materials formulieren.

Bisher folgt auf die Suche nach "Heinrich", welcher der meistgesuchte Begriff in den RI ist, die Anzeige[2]:

Sie suchten nach 'Heinrich' Ihre Suche erzielte 17101 Treffer, ausgewählt wurden die 1000 relevantesten Regesten. Sie sehen die Treffer 1 bis 20. Zur Verfeinerung Ihres Ergebnisses modifizieren Sie Ihre Suchabfrage.

Man könnte die Suchmöglichkeiten vielleicht dahingehend ergänzen, dass dem Nutzer bei der Suche nach 'Heinrich' folgendes angeboten wird:

"Sie suchten nach 'Heinrich' Sie haben 17.101 Treffer. Möchten Sie für die Einschränkung der Treffermenge eine Visualisierung der Trefferliste in chronologischer oder geographischer Form oder die Anzeige der Treffer pro Abteilung der Regesta Imperii ?"

Mit neuen Visualisierungsmethoden und einem transparenten Drill-Down[3] könnten neue Blicke auf bereits vorhandenes digitales Material möglich werden.

4 Die kritische Masse

In den letzten Jahren haben die als digitale Volltexte zur Verfügung stehenden Quellen stark zugenommen. Neben den Regesta Imperii werden im Akademienprogramm[4] immer mehr Projekte digitalisiert und im Netz bereitgestellt. Bei Neu-

[2] Vgl. http://www.regesta-imperii.de/regesten/suche.html abgerufen am 10.10.2014.
[3] https://de.wikipedia.org/w/index.php?title=Drill-Down&oldid=117104292
[4] Zum Akademienprogramm vgl. http://www.akademienunion.de/forschung/

anträgen im Akademienprogramm muss ein Abschnitt zur Bereitstellung der Forschungsergebnisse im Internet enthalten sein. Diese Bemühungen für eine breite Digitalisierung von Forschungsmaterialien haben in den letzten Jahren dazu geführt, dass wir langsam eine "kritische Masse" überschritten haben[5]. Und hier würde ich wieder zum "Google-Anspruch" aus dem letzten Absatz zurückkehren. Könnte es nicht neue Forschungsperspektiven aufzeigen, wenn wir große Datensammlungen gemeinsam untersuchen, sehr große Ergebnismengen erhalten und aus den anschließenden Visualisierungen oder mit anschließendem Drill-Down neuen Phänomenen oder Fragestellungen auf die Spur kommen, die wir aus analoger Perspektive nicht wahrgenommen haben ?

5 Visualisierung als Weg zu neuen Erkenntnissen

In meinem Beitrag zu Digitalen Perspektiven mediävistischer Quellenrecherche habe ich verschiedene Suchmasken von Online-Quellenportalen untersucht. Dabei konnte ich zeigen, dass die Suchmasken in der Regel optimale Möglichkeiten für die Einschränkung der Treffermenge auf eine zu handhabende Größe bieten. Dem gegenüber werden bei der Trefferanzeige kaum Möglichkeiten zur Weiterverarbeitung oder Visualisierung von großen Ergebnismengen geboten. Gerade hier liegt aber aus meiner Sicht eine große Chance für die Geschichteswissenschaften: die Untersuchung großer Quellenbestände im Sinne einer Big Data History, mit der Zusammenhänge aufgezeigt werden können, die im "analogen" Zeitalter nicht möglich waren.

6 Fazit

Momentan stehen wir noch an der Schwelle zu Big Data History. Es ist aber nur noch eine Frage der Zeit, bis gemeinsame Schnittstellen projektübergreifende

[5]Ein Hinweis, dass wir die "kritische Masse" überschritten haben, war der Erfolg von http://codingdavinci.de/

Quellenanalyse möglich machen, die Ergebnisse visualisiert und weiterverarbeitet werden können und damit neue Blicke auf das Quellenmaterial möglich werden. Die Analyse großer Datenmengen bringt für den Historiker aber auch Herausforderungen mit sich. Bei großen Datenmengen stellt sich die Frage nach Fehlerabschätzungen, neuen Analysemethoden und theoretischen Ansätzen. Andererseits verspricht die Digitale Perspektive auf eine Big Data History interessante neue Blicke auf unser Quellenmaterial.