

# Urkunden als Netzwerk

Ein Werkstattbericht von Daniel Jeller<sup>1</sup>

Mit Urkunden, also schriftlichen Aufzeichnungen von Rechtshandlungen, die besonders im Mittelalter und der Frühen Neuzeit in Gebrauch waren<sup>2</sup>, verfügt die Geschichtswissenschaft über eine der wohl vielfältigsten und spannendsten Quellengattungen. Für die Erschließung und Erforschung dieser Dokumente waren dabei traditionell neben den Originalen gedruckte Editionen der Urkunden das Hauptmedium. Im Digitalen Zeitalter rückt allerdings vermehrt die Benutzung von "digitalen" Editionen auf Datenträgern und in online-Datenbanken in den Fokus. Neben dem praktischen Vorteil der leichten und stetigen Verfügbarkeit in diesem Medium birgt die Publikation von Urkunden online in digitaler Form weitere Möglichkeiten für Forscherinnen und Forscher, die sich mit Urkunden beschäftigen.<sup>3</sup>

Ihrer Entstehung in mittelalterlichen und frühneuzeitlichen Institutionen nach sind Urkunden inhaltlich stark vernetzte Dokumente, in denen vielfältige Akteure wie die päpstliche Kanzlei, geistliche und weltliche Würdenträger oder Grundbesitzer miteinander in Verbindung treten. Eine Darstellung dieses teilweise sehr umfangreichen Netzwerks aus Personen, Institutionen und Orten in Zeit und Raum ist in der Form einer klassischen Edition (egal ob gedruckt oder in digitaler Form) nur schwer möglich. Erste Abhilfe bieten Urkundendatenbanken wie die des Monasterium-Portals<sup>4</sup>, indem Sie auf übersichtliche Weise einen Überblick über die beinhaltenden Daten bieten. Dabei bleibt jedoch meist die aus dem Buchwesen vertraute Form eines durch Inhaltsverzeichnis oder Index erschlossenen Textes bestehen und der Netzwerkcharakter bleibt dem Benutzer weitgehend verborgen. Diesem Umstand verschaffen auch die häufig sehr komplexen Suchmöglichkeiten innerhalb der Datenbanken und Dokumente nur bedingt Abhilfe.

Die Informatik hat jedoch in den letzten Jahrzehnten einen neuen Datenbank-Typus entwickelt, der auf der aus der Mathematik bekannten Graphentheorie aufbaut und das Speichern und Abfragen von Informationen in Form eines Netzwerks aus benachbarten Knoten und ihren Verbindungen ermöglicht. Die so gespeicherten Daten lassen sich mit geeigneten Algorithmen durchsuchen und neben tabellarischen oder textuellen Exzerpten auch in diagrammatischer Form darstellen und bieten so zahlreiche neue Möglichkeiten zur Suche in den Inhalten und Visualisierung von Zusammenhängen. Um die *klassischen* digitalen Urkundeneditionen für dieses Medium aufzubereiten müssen die Inhalte einer Urkunde ihrem semantischen Gehalt nach analysiert und in ein Modell der vernetzten Struktur in zeitlichem, räumlichen und personellen Kontext transformiert werden (siehe Abbildung 1).

Der vorliegende Beitrag versteht sich als erster Versuch, diese Schritte an Urkunden aus der Monasterium-Datenbank zu vollziehen und die dabei auftretenden Probleme und Lösungsansätze zu dokumentieren. Zudem soll, in Form einer einfachen Beispielanwendung, ein Ausblick auf die Möglichkeiten, die eine derartig aufbereitete Urkundendatenbank der modernen Diplomatik bietet, gegeben werden.

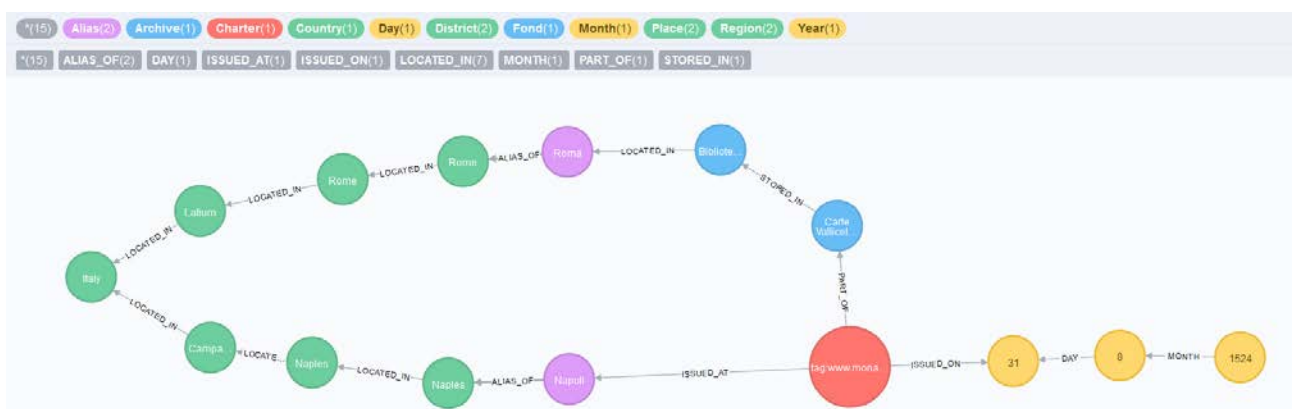


Abbildung 1: Eine Urkunde (rot) in ihrem geographischen (grün und lila) und zeitlichen Kontext (gelb), dargestellt als gerichteter Graph. Abb. des Autors.

## Das Monasterium-Portal aus datentechnischer Sicht

Mit seinen aktuell etwas über sechshunderttausend Urkunden<sup>5</sup> aus einer Vielzahl von europäischen Archiven ist das Monasterium-Portal und sein online Archiv *MOM-CA*<sup>6</sup> eine der größten frei zugänglichen Sammlungen von mittelalterlichen und frühneuzeitlichen Urkunden in Bild und Text. Dabei unterscheiden sich die einzelnen Urkunden in ihrer Erschließungstiefe teilweise stark voneinander da die teilnehmenden Archive direkt verantwortlich für die zur Verfügung gestellten Daten sind. Je nach Zeitbereich, geografischer und historischer Relevanz der Dokumente sowie länderspezifischen Erschließungsstandards liegen die Metadaten der Digitalisate meist zwischen folgenden Polen:

1. Minimaldaten in Form von Archiv-Signatur, Datierung, häufig mit einem Regest
2. Hohe Erschließungstiefe mit Signatur, Datum, Regest, Volltext, diplomatischem Apparat, Personen-, Orts- und Sachindex

Die vorhandenen Metadaten, damit sind in Folge alle in der Monasterium-Datenbank gespeicherten und damit abrufbaren Informationen zu einer Urkunde, also Regest, Volltext, Index, etc. gemeint, sind dabei wie bereits in obengenannten Punkten angedeutet, nicht nur in ihrem Vorhandensein, sondern auch in ihrer spezifischen Ausprägung äußerst inhomogen. Die Kodierung dieser Daten mithilfe des eigens für die Erschließung von Urkunden entwickelten *CEI-XML-Schemas*<sup>7</sup> hat zwar zur Folge, dass die verfügbaren Daten formal stark standardisiert und daher gut zur Weiterverarbeitung geeignet sind, die vielfältige Herkunft der Daten sowie die Möglichkeit für Benutzer, im ebenfalls frei zugänglichen<sup>8</sup>, grafischen Editor *EditMOM* die Metadaten einer Urkunde, ähnlich den Einträgen in der Wikipedia, zu ändern bedingt jedoch, dass nur in geringem Maße universal gültige Inhaltsstandards angelegt werden können. Dies erschwert die Verwendung der Daten. Um detaillierte Auswertungen vornehmen zu können müssen die Informationen daher aufbereitet werden.

## Versuch: Daten aus MOM-CA in neuem Kontext

Um konkrete Erfahrungen in der Arbeit mit Urkunden in einer graphenbasierten Struktur zu sammeln ist es sinnvoll, nicht nur theoretische Überlegungen anzustellen, sondern ein konkretes Testprojekt zu entwerfen. Dafür wurde folgende Funktionalitäten vorgesehen:

Eine Browser-basierte Applikation sollte entwickelt werden, die auf der Graphendatenbank *Neo4j*<sup>9</sup> aufbauend Urkunden aus ausgewählten Beständen<sup>10</sup> auf einer Karte darstellen sollte. Dabei sollte es möglich sein, die Darstellung auf beliebige Zeitbereiche zu beschränken. Einzelne Urkunden sollten dabei nicht direkt auf der Karte aufscheinen, sondern durch die in ihnen erwähnten Orte und Personen repräsentiert werden. Beim Klick auf einen derartigen Inhalt sollte eine Zusammenfassung der zugrundeliegenden Urkunde angezeigt werden.<sup>11</sup>

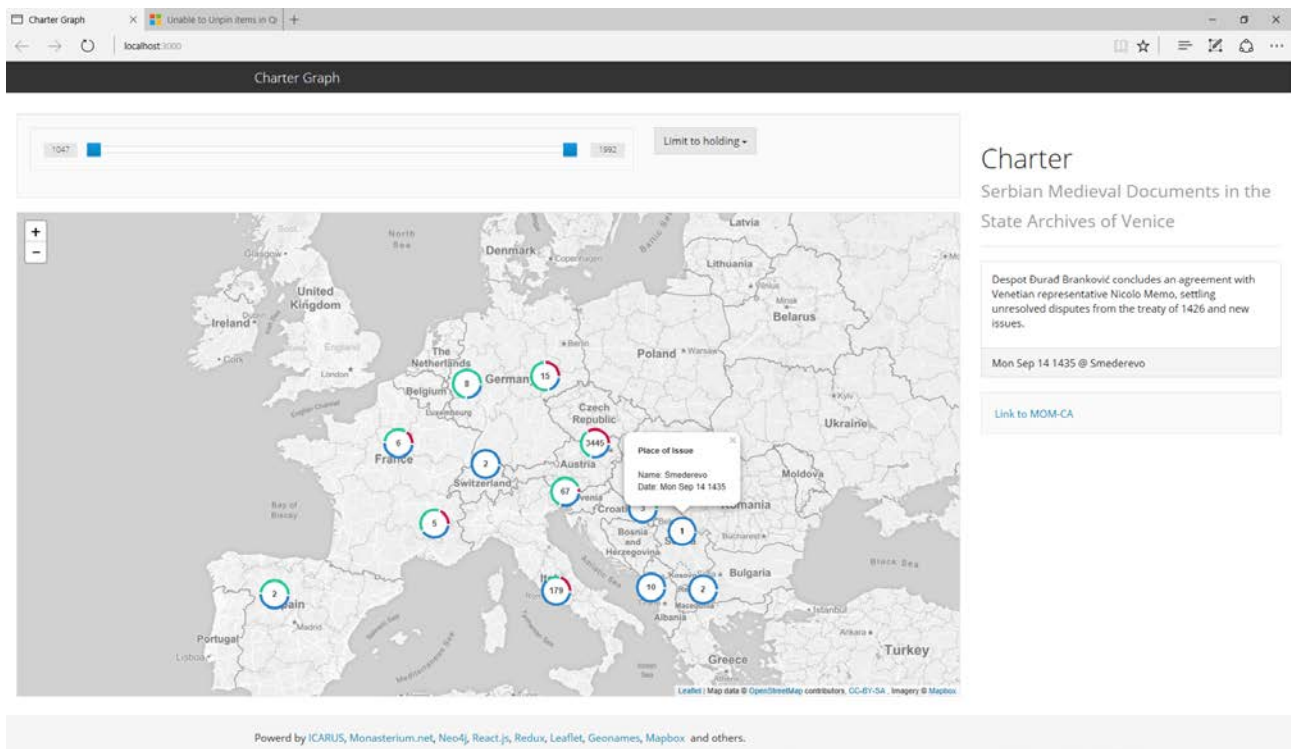


Abbildung 2: Screenshot der fertigen Applikation. Abb. des Autors.

Um diese Applikation umsetzen zu können waren folgende Arbeitsschritte nötig:

1. Entwurf der benötigten Datenstruktur auf Graphenbasis
2. Export der Daten aus Monasterium
3. Konversion der Daten in die neue Datenstruktur
4. Speichern in der Datenbank
5. Entwicklung der Applikation

Für die geplante Applikation musste dabei auf folgende Metadaten Rücksicht genommen werden:

- Archivspezifischer Kontext (Archiv, Bestand, Signatur)
- Datum und Regest der Urkunden
- Ausstellungsdatum und -ort sowie Aussteller
- Sonstige erwähnte Orte und Personen

Diese Daten in eine einfache Graphenstruktur (siehe Abbildung 3) überführt werden. Dazu wurden die XML Strukturen mithilfe eines Java Programms umgewandelt, die Orts- und Personennennungen normalisiert und identifiziert und alles in eine bereitgestellte Neo4j-Datenbank geschrieben. Diese Daten bildeten dann die Grundlage für die fertige JavaScript-Anwendung.

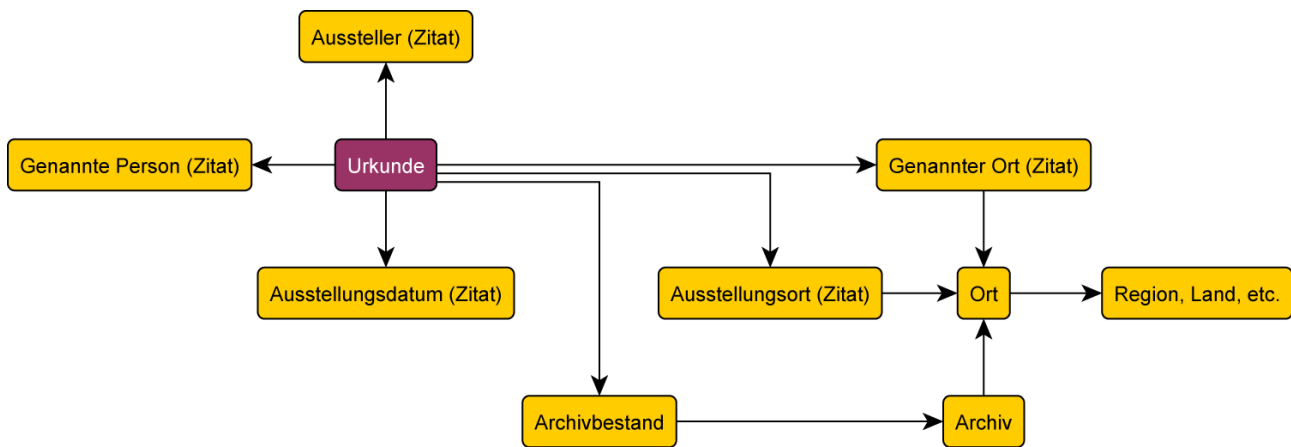


Abbildung 3: Schematische Darstellung einer Urkunde als Graph. Abb. des Autors.

## Daten in Monasterium

Wie bereits erwähnt werden die Metadaten der Urkunden in Monasterium im CEI-XML Schema ausgezeichnet und in der internen Datenbank gespeichert. Für Personen und Orte sieht das Schema im Allgemeinen folgende Struktur vor:

1. Der Originaltext, also beispielsweise *Carolus* im XML-Element-Text. Meistens ein Zitat.
2. Eine vom jeweiligen Bearbeiter frei gewählte Form des Zitats im `reg`-Attribut, die im Optimalfall einer normalisierten Form des Textes entspricht, also *Karl der Große*, aber auch *Carlo Magno* (auf Italienisch) oder *I. Károly frank császár* und *I. Nagy Károly* (beides auf Ungarisch). Es findet hier keine weitere Normierung durch das System statt.

Die exakte Schreibweise (*Carolus* oder *Karolus*), also was alles als Originaltext/-name ausgezeichnet wird und eine etwaige Normalisierung bleiben letztlich völlig dem Bearbeiter überlassen. Dieser ist auch nicht dazu verpflichtet, das `reg`-Attribut überhaupt zu verwenden. Zudem können Namen auch unabhängig von einer Transkription des Textes in Form eines eigens angelegten Indexes ausgezeichnet werden. Hier kann auch der Text-Inhalt (Punkt #1 in der obigen Liste) bereits eine normalisierte Version eines Namens sein. Es können also beispielsweise folgende XML-Fragmente gleichberechtigt in der Datenbank koexistieren. Alle beziehen sich auf Karl den Großen:

1. `<cei:persName>Karolus</cei:persName>`
2. `<cei:persName>Karl der Große</cei:persName>`
3. `<cei:persName reg="Karl der Große">Carolus</cei:persName>`
4. `<cei:persName reg="Karl (I.), der Große">Carolus</cei:persName>`
5. `<cei:persName reg="Carlo Magno">Karolus</cei:persName>`
6. `<cei:persName reg="I. Károly frank császár">Karolus</cei:persName>`

Im Gegensatz dazu bezeichnen folgende Fragmente Kaiser Karl IV. (\* 1316; † 1378):

1. `<cei:persName>Karolus</cei:persName>`
2. `<cei:persName>Karl IV.</cei:persName>`
3. `<cei:persName reg="Karl (IV.)">Karolus</cei:persName>`
4. `<cei:persName reg="Karel, císař římský">Karl</cei:persName>`

Das Problem springt ins Auge: ohne Berücksichtigung des Kontextes sind die jeweils ersten Nennungen von Karl dem Großen und Karl IV. nicht voneinander zu unterscheiden. Zudem ist es äußerst schwer, automatisiert festzustellen, dass *Karl der Große* identisch mit *I. Károly frank császár* ist, nicht aber *Karel, císař římský*. Dasselbe gilt in noch größerem Ausmaß für Namensnennungen von weniger bedeutenden Personen und natürlich auch für Ortsnamen. Folgendes sind beispielsweise real in der Datenbank existierende Einträge:

- `<cei:persName reg="Marian Zwinger, Abt von Melk (1819-1837)">Marian von Melk</cei:persName>`
- `<cei:persName reg="Christoph Barkmann, Bewohner von Lusdorf">Christoph Barkmann</cei:persName>`
- `<cei:placeName>Geras</cei:placeName>`
- `<cei:placeName>Stift Geras</cei:placeName>`
- `<cei:placeName reg="Geras, Prämonstratenserabtei">Jeros</cei:placeName>`
- `<cei:placeName reg="Geras, Prämonstratenserabtei">Geras</cei:placeName>`
- `<cei:placeName reg="Stift Geras">Geras</cei:placeName>`
- `<cei:placeName reg="Liebenberg (Gde. Ludweis-Aigen, GB WT)">Liebenberg</cei:placeName>`
- `<cei:placeName reg="Reinthal (GB LT) (?)">Reyntal</cei:placeName>`
- `<cei:placeName reg="Ranzern (Rancirov, MÄ/CZ, GB Jihlava)">Ranzern</cei:placeName>`
- `<cei:placeName>pod Krakowem</cei:placeName>`

Um es nun einer Anwendung zu ermöglichen, diese sehr vielfältigen Namen zueinander in Beziehung zu setzen, ist es nötig, eine Möglichkeit zu finden, diese Beziehungen mit ausreichender Sicherheit festzustellen.

## Identifikation von Orten

Wie im vorigen Abschnitt geschildert, verfügt Monasterium für Personen und Orte lediglich über textuelle Informationen die aus dem Originaltext der Urkunden abgeleitet werden. Um die Orte auf der Landkarte darstellen zu können müssen also zu diesen Ortsnennungen die entsprechenden realen Orte beziehungsweise deren Koordinaten ermittelt werden. Als Hilfsmittel dazu wurde die Ortsdatenbank *GeoNames*<sup>12</sup> verwendet. Dort können mithilfe einer frei zugänglichen Schnittstelle Orte mithilfe ihres Namens gesucht und ihre geographische Einbettung in Bezirke, Regionen und Länder, ihre heutigen Koordinaten sowie eine Liste von Namen der Orte in verschiedenen Sprachen abgerufen werden. Das Hauptproblem ist hier, die teils sehr komplexen Ortsnennungen in Monasterium (z.B. *Liebenberg (Gde. Ludweis-Aigen, GB WT)*) soweit zu normalisieren, dass der betreffende Ort automatisch in der Datenbank gefunden werden kann sowie soweit als möglich Fehlidentifikationen zu vermeiden. Dabei ist vor allem die Tatsache von Nachteil, dass im internationalen aber auch regionalen Kontext viele Orte ähnliche oder gleiche Namen in ihrer oder einer anderen Landessprache haben. Im vorliegenden Fall wurde auf folgende Arten versucht, mit diesem Problem umzugehen:

1. Die Namen wurden schrittweise normalisiert (Entfernung von Sonderzeichen) und in einzelne Zeichenketten getrennt (aus *Liebenberg (Gde. Ludweis-Aigen, GB WT)* wird hier "Liebenberg", "Gde. Ludweis-Aigen", "GB WT", "Gde.", "Ludweis-Aigen", "Ludweis", "Aigen"). Diese Begriffe wurden dann nach gewissen Kriterien gereiht (komplexe Begriffe zuerst, kurze zuletzt) und in der Datenbank gesucht. Die Ergebnisse wurden dann auf ihre Richtigkeit überprüft.

2. Die Suche wurde manuell auf Europa beschränkt, um Fehlidentifikationen von Orten vor allem in Amerika auszuschließen, da die Orte dort oft denselben Namen wie europäische haben können und oft durch ihre Größe von der GeoNames-Suche bevorzugt geliefert werden. Für die Testbestände aus Mitteleuropa können diese Orte aber ausgeschlossen werden.
3. Um Fehlidentifikationen innerhalb Europas weitgehend ausschließen zu können wurde der geographische Kontext einer Urkunde ermittelt und mitberücksichtigt. So wurde zuerst immer der Ausstellungsort der Urkunde untersucht, und dann alle anderen Orte in derselben Urkunde erwähnten Orte im räumlichen Bezug auf den Ausstellungsort ermittelt. (siehe Abbildung 4)
4. Dem Autor zugängliche zusätzliche Informationen zu den Beständen wurden berücksichtigt. So war beispielsweise bekannt, dass der Bearbeiter der Urkunden des Stiftes Geras soweit als möglich im *reg*-Attribut auch den Gerichtsbezirk in Abkürzung vermerkt hatte. *GB WT* steht beispielsweise für *Gerichtsbezirk Waidhofen an der Thaya*. Dies stimmt häufig mit den in GeoNames zu Orten gespeicherten administrativen Einheiten überein, weshalb Treffer außerhalb dieser Einheit auch weitgehend ausgeschlossen werden können.

Konnte mit diesem Vorgehen kein passender Ort gefunden werden, etwa, weil der Ort heute nicht mehr existiert oder weil der Name nicht mit ausreichender Sicherheit identifiziert werden konnte, so wurde der betreffende Ort für die Darstellung ignoriert.

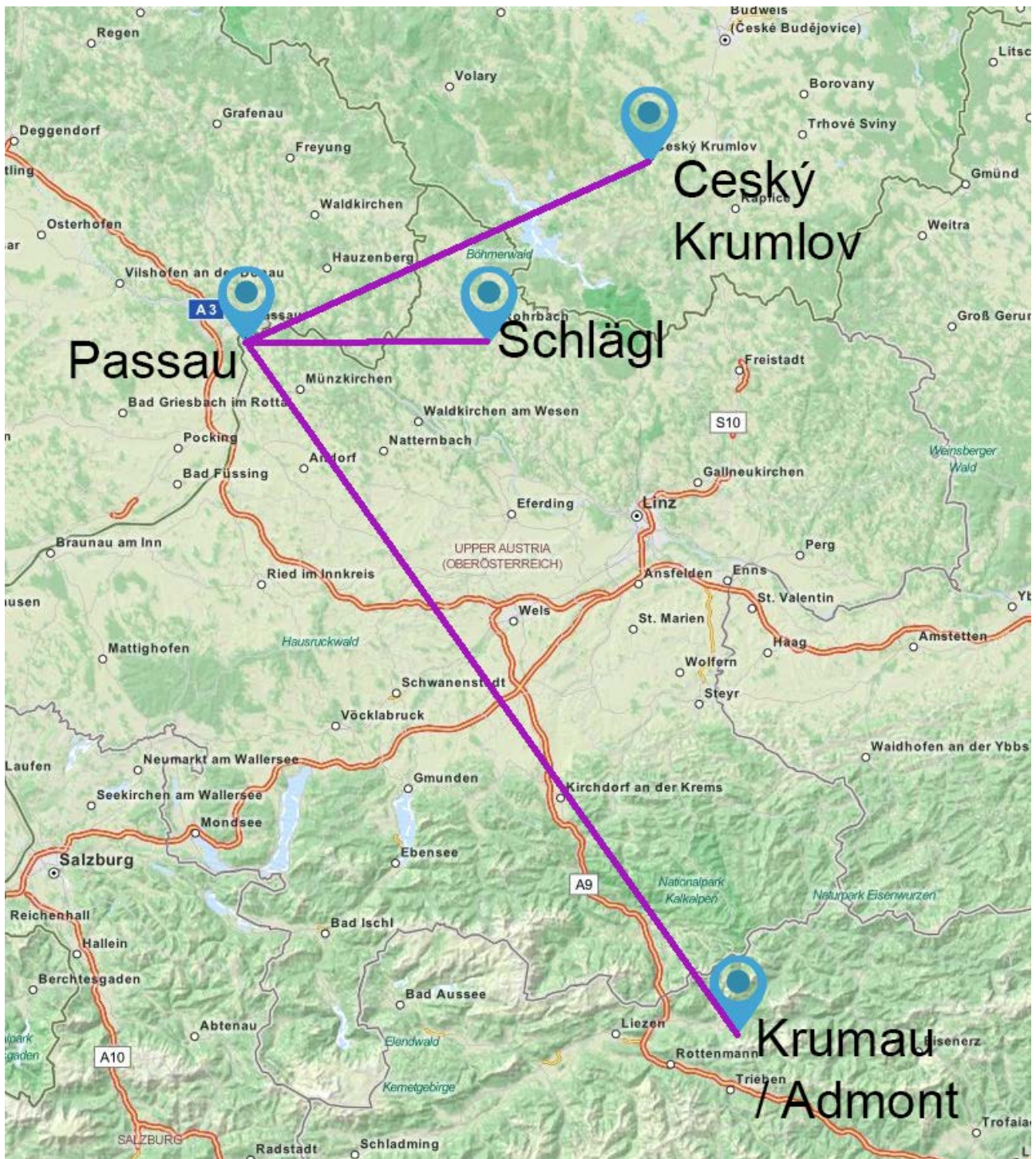


Abbildung 4: Schematische Darstellung der Identifikation des Ortes Český Krumlov (auf Deutsch Krumau) in einer Urkunde mit dem Ausstellungsort Passau aus dem Stiftsarchiv Schlägl. Aufgrund der geographischen Lage von Passau und Schlägl scheint es wahrscheinlicher, dass mit "Krumau" Český Krumlov in Tschechien und nicht Krumau bei Admont gemeint ist.

## Probleme

Mit der im vorigen Abschnitt skizzierten Methode konnten den 1376 in den Urkunden erwähnten Orten 452 real existierende Orte zugeordnet werden. Um festzustellen, wie exakt die Methode ist, wurden die 368 für Geras identifizierten Orte manuell auf ihre Richtigkeit überprüft und dabei mindestens 42 Fehlidentifikationen gefunden. Damit beträgt die Fehlerquote in etwa 10%, wobei die Urkunden von Geras allerdings einen Optimalfall darstellen. Wie bereits erwähnt wurde war hier der Suchmechanismus nicht nur auf die reinen Textdaten angewiesen, sondern konnte auch die Information der Gerichtsbezirke zur Kontrolle

verwenden. Für Bestände für die keine derartigen Zusatzinformationen vorliegen ist die Fehlerquote wohl deutlich höher anzusetzen.

## Identifikation von Personen

Noch schwieriger ist die Identifikation von Personen. Lässt sich für bedeutende Personen wie Päpste oder Kaiser sicherlich auch eine Online-Quelle finden, so ist die Situation für Personen der Regionalgeschichte wesentlich unvorteilhafter. Aus Gründen der Einfachheit wurde daher für die vorliegende Anwendung lediglich eine Identifikation aufgrund eines einfachen Vergleichs der Namen angestellt.

## Fazit

Nachdem die Daten in der beschriebenen Weise aufbereitet und in die Graphendatenbank eingespielt waren, wurde die in Abbildung 2 dargestellte Anwendung in JavaScript umgesetzt. Die einzelnen Orte wurden aus der Datenbank eingelesen und auf der Karte mithilfe der in GeoNames hinterlegten Koordinaten dargestellt. Die identifizierten Personen wurden am gleichen Ort dargestellt, an dem die Urkunde ausgestellt wurde. In der praktischen Benutzung zeigt sich, dass eine derartige Darstellung etwa einen intuitiven Zugang zum Einflussbereich eines Klosters im Laufe seiner Geschichte ermöglichen kann (siehe Abbildung 5). Auch einige Identifikationsfehler lassen sich in der Darstellung schnell ausmachen. Beispielsweise springt einem in Abbildung 5 schnell die einzelne Nennung eines Ortes in Spanien ins Auge. Bei näherer Betrachtung stellt man fest, dass hier ebenfalls ein Ort namens Geras liegt (siehe Abbildung 6). Ein Blick auf die Details zur Nennung zeigt als Ausstellungsort der betreffenden Urkunde Bordeaux in Frankreich. Nach der im Kapitel *Identifikation von Orten* geschilderten Vorgehensweise wurde vom System, ausgehend von Bordeaux, Geras in Spanien als der wahrscheinlichste Ort festgestellt. Der Blick auf den Kontext des Stiftsarchivs in Geras lässt dies natürlich sofort als Fehlidentifikation erkennbar sein. Hierin liegt auch der naheliegendste Faktor für eine eventuelle Weiterentwicklung des Identifikationsprozesses:

Es reicht nicht, wie im vorliegenden Fall jede Urkunde isoliert zu betrachten. Es wäre angebracht, nach dem ersten Identifikationsprozess weitere Durchläufe vorzunehmen. In diesen könnte die Wahrscheinlichkeit, dass eine Identifikation eines Ortes korrekt ist im Hinblick auf alle bereits identifizierten Orte einer relativ eng gefassten Gruppe von Urkunden beurteilt werden. Eine solche Einheit könnten etwa alle Urkunden eines Bestandes in einem definierten Zeitraum, etwa 1200 bis 1250 sein. Dies würde beispielsweise Geras in Spanien für diesen Zeitraum sehr unwahrscheinlich machen, da alle anderen Nennungen von *Geras* innerhalb der Einheit mit Geras in Niederösterreich identifiziert wurden.

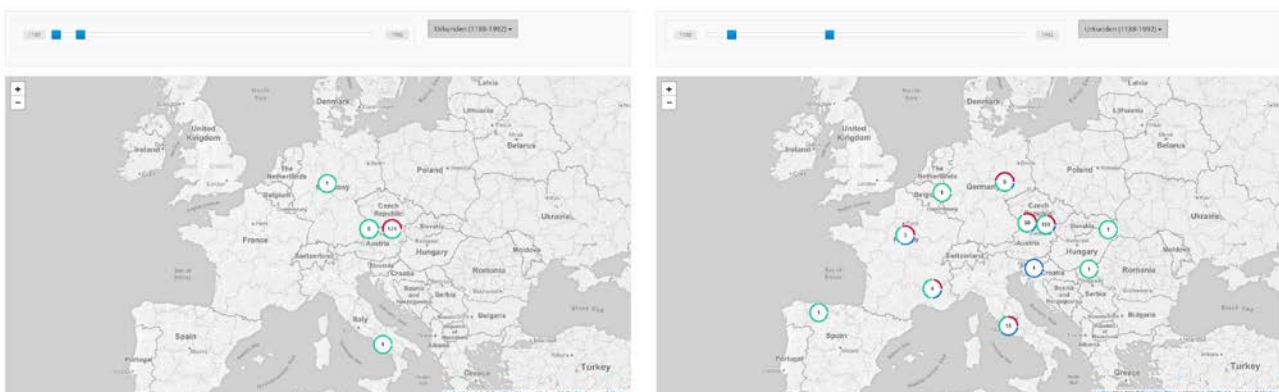


Abbildung 1: Orte (Ausstellungen in Blau, sonstige Nennungen in Grün) und Personen (Rot) in Urkunden im Stiftsarchiv Geras zwischen 1188 und 1250 (links) und zwischen 1250 und 1500 (rechts). Abb. des Autors.



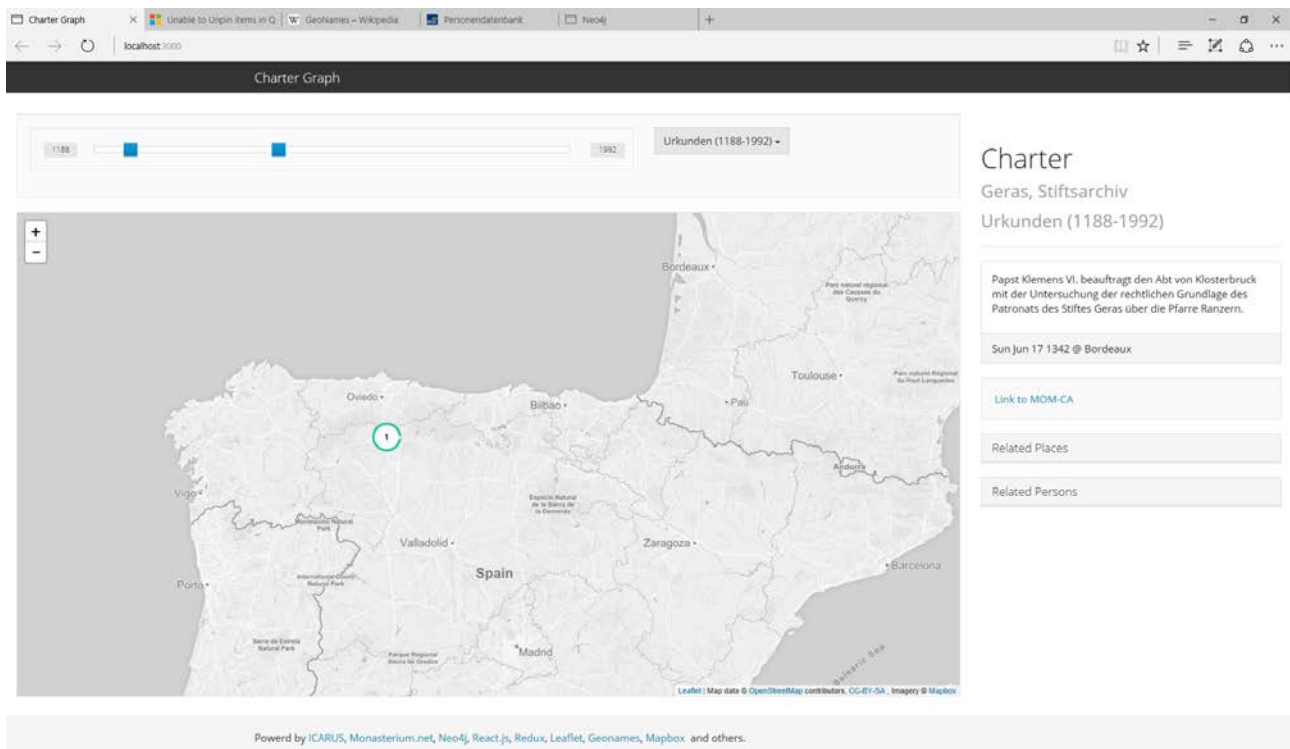


Abbildung 2: Geras in Spanien. Abb. des Autors.

Zusammenfassend lässt sich sagen, dass die eine Aufbereitung von Urkunden in Form in einer Graphenstruktur, wie im vorliegenden Text geschildert, großes Potenzial bietet, da die darauf aufbauenden Anwendungen stark expressiv sein können und intuitivere Zugänge ermöglichen können als traditionelle Darstellungsformen. Der konkrete Versuch hat allerdings auch gezeigt, dass noch einiges an Vorarbeiten zu leisten sein wird, bevor derartige Transformationen zwischen unterschiedlichen Datenstrukturen mit ausreichend hoher Sicherheit automatisch vorgenommen werden können. Es lohnt sich aber auf jeden Fall, in diesem Bereich weitere Arbeiten anzuschließen.

<sup>1</sup> Dieser Text ist unter <http://documents.icar-us.eu/documents/2016/08/urkunden-als-netzwerk.pdf> beziehbar.

<sup>2</sup> Josef Hartmann, Urkunden, in: Friedrich Beck und Eckart Henning (Hg.), Die archivalischen Quellen, Köln and Wien [u.a.] 2012, 25–54.

<sup>3</sup> Siehe dazu u.a. die Diplomarbeit des Autors: Daniel Jeller, Die Archivalie im Zeitalter ihrer digitalen Reproduzierbarkeit, Unveröffentlichte Diplomarbeit, Universität Wien 2013. online unter: [http://clus-ter.nettek.at/?post\\_type=document&p=1118](http://clus-ter.nettek.at/?post_type=document&p=1118) (01.8.2016).

<sup>4</sup> <http://monasterium.net>

<sup>5</sup> ICARUS, Home - Monasterium.net, online unter <http://monasterium.net/mom/home> (2.8.2016).

<sup>6</sup> Abgekürzt für Monasterium-Collaborative Archive.

<sup>7</sup> Charters Encoding Initiative, CEI - The Project, online unter <http://www.cei.uni-muenchen.de/project> (2.8.2016)

<sup>8</sup> Lediglich eine Anmeldung im System ist nötig, um inhaltliche Änderungen moderieren zu können und somit die Qualität der bearbeiteten Daten sicherzustellen.

<sup>9</sup> <https://neo4j.com>

<sup>10</sup> Ausgewählt wurden 5 Bestände mit insgesamt 1722 Urkunden.

<sup>11</sup> Siehe Abbildung 2.

---

<sup>12</sup> <http://www.geonames.org>