Gradl, Tobias; Henrich, Andreas

Nutzung und Kombination von Daten aus strukturierten und unstrukturierten Quellen zur Identifikation transnationaler Lebensläufe

Datum der Zweitveröffentlichung: 13.03.2025

Verlagsversion (Version of Record), Konferenzveröffentlichung

Persistenter Identifikator: urn:nbn:de:bvb:473-irb-1070103

Erstveröffentlichung

Gradl, T.; Henrich, A. (2017): Nutzung und Kombination von Daten aus strukturierten und unstrukturierten Quellen zur Identifikation transnationaler Lebensläufe, in: E. Burr (Hrsg.), DHd 2016 : Modellierung, Vernetzung, Visualisierung : die Digital Humanities als fächerübergreifendes Forschungsparadigma : Konferenzabstracts : Universität Leipzig 7.-12. März 2016, 2. überarb.und erw. Ausg., Duisburg: nisaba, S. 261–265, doi: 10.5281/zenodo.3679331.

Nutzung und Kombination von Daten aus strukturierten und unstrukturierten Quellen zur Identifikation transnationaler Lebensläufe

Gradl, Tobias
tobias.gradl@uni-bamberg.de
Universität Bamberg, Deutschland

Henrich, Andreas
andreas.henrich@uni-bamberg.de
Universität Bamberg, Deutschland

Biographien erscheinen als nahezu ubiquitärer Forschungsgegenstand in den unterschiedlichsten geisteswissenschaftlichen Disziplinen. Neben der qualitativen Betrachtung wurden aus diesem Grund auch Verfahren zur quantitativen Analyse biographischer Texte entwickelt, die zumeist die Identifikation und Extraktion relevanter Merkmale aus unstrukturiertem Text behandeln. So präsentieren beispielsweise Bamman und Smith eine Methode zur unüberwachten Erkennung biographischer Daten in unstrukturiertem Text (Bamman / Smith 2014). Blessing und Kuhn präsentieren mit ihrem Konzept und webbasiertem Prototypen zur Emigrationsanalyse eine konkrete Anwendung der quantitativen Analyse und Aggregation biographischer Daten (Blessing / Kuhn 2014).

Auf Basis der Machbarkeitsstudie *»Cosmobilities« – Grenzüberschreitende Lebensläufe in den europäischen Nationalbiographien des 19. Jahrhunderts*

motivieren wir in diesem Vortrag die Notwendigkeit einer kombiniert qualitativen und quantitativen Betrachtung unterschiedlicher Quellen biographischer Daten – einer Aufgabe, der nach unserer Einschätzung aktuell eher wenig Priorität zugewiesen wird. Als Schwerpunkte vertiefen wir anschließend zwei für die Korrelation und Integration relevanter Daten wichtige Aspekte: Zum einen wird mit der *kontextspezifischen Kombination biographischer Daten* ein *iterativer Ansatz* vorgestellt, der bei der Verknüpfung von Einzelergebnissen der quantitativen Verfahren unterstützt und die Berücksichtigung qualitativer Resultate zulässt. Der zweite Schwerpunkt des Vortrags fokussiert auf die *Unterstützung des für die Erstellung biographischer Profile notwendigen Verarbeitungsprozesses* durch Komponenten der DARIAH-DE Infrastruktur, welche die Erweiterung des Prototypen um zusätzliche strukturierte und unstrukturierte Datenquellen erleichtern.

Fachwissenschaftlicher Kontext

Historiker des Leibniz-Instituts für Europäische Geschichte Mainz und der Ludwig-Maximilians-Universität München untersuchten im Rahmen von *Cosmobilities* exemplarisch, inwiefern biographische Texte über transnationale Bezüge einer Person hinwegtäuschen. Eine Besonderheit der Transnationalität [1] besteht darin, dass sich diese oft erst durch Betrachtung unterschiedlicher Quellen als solche zu erkennen gibt: Durch ihre nationale Prägung beschreiben biographische Texte – insbesondere in den Nationalbiographien – eine Person aus einer nationalen Perspektive und vernachlässigen oder verschweigen Einflüsse der Person auf andere Nationen oder Kulturkreise.

Transnationalität in Lebensläufen: Ein Beispiel

Betrachten wir als Beispiel den 1847 geborenen, jüdischen Bankier Jakob Heinrich Schiff. Nach Geburt und Kindheit in Frankfurt migrierte dieser zunächst im Alter von 18 Jahren und – nach drei Jahren in Hamburg und Frankfurt – 1875 ein weiteres Mal in die USA.

Der rund 950 Wörter umfassende Eintrag zu Jakob Schiff in der deutschsprachigen Wikipedia gibt Aufschluss über die Transnationalität in seinem Leben und betont insbesondere auch berufliche Stationen als Bankier. Der mit rund 2.350 Wörtern umfassendere, englischsprachige Artikel unterscheidet sich vor allem durch die differenzierte Betrachtung des Philanthropen und Geschäftsmanns und seine weitreichende finanzielle Unterstützung Japans im Krieg gegen Russland 1904-1905. Obwohl beide Artikel jeweils die wesentlichen Aspekte seines Lebens umfassen, enthalten diese auch Informationen, die dem jeweils Anderen fehlen: So erwähnt nur der deutsche Eintrag Schiffs Brüder und beschreibt seine Rolle als Gründungsmitglied der Johann Wolfgang Goethe-Universität. Im englischsprachigen Beitrag fehlen diese Informationen, während aber eine detaillierte Auflistung der von ihm unterstützten, in den Vereinigten Staaten ansässigen Einrichtungen vorgelegt wird.

Historiker können für eine fundierte Auseinandersetzung mit dem Leben von Jakob Schiff auf einen Eintrag der Datenbank von *Immigrant Entrepreneurship* zurückgreifen. In dieser führt das 1987 gegründete Deutsche Historische Institut Washington (DHI) fundierte, redaktionell geprüfte Einträge zu Deutsch-Amerikanischen Unternehmern. Schiff ist dort mit einem über 10.000 Worte umfassenden Artikel verzeichnet. Und obwohl der Artikel eine historisch differenzierte Analyse seines Lebens und Wirkens liefert: einige in der Wikipedia verfügbare Informationen (z. B. Informationen über die Brüder und seine Stiftung des orientalischen Seminars an der Universität Frankfurt) fehlen auch hier.

Wikipedia als biographische Quelle

Das Beispiel Jakob Schiffs erlaubt zwei direkte Rückschlüsse: Erstens, dass oft erst durch die Kombination nationaler Perspektiven ein übergreifender Eindruck über eine transnationale Biographie entstehen kann. Zweitens kann die Wikipedia zwar aufgrund ihrer Intention und Ideologie nicht als Quelle historischer Forschung dienen; für die Identifikation und initiale Analyse der Transnationalität von Biographien bietet die Wikipedia jedoch den Vorteil einer – insbesondere gegenüber den Nationalbiographien – oft weitaus geringeren nationalen Prägung. Vor allem jedoch stehen Wikipedia-Artikel in den verschiedensten Sprachen frei und ohne Zugriffshürden zur Verfügung, worin ein bedeutender Vorteil für die Anwendung quantitativer Verfahren liegt: Allein die deutschsprachige Wikipedia beinhaltet etwa 560.000 Einträge zu Personen. In Kontrast hierzu stellen die ebenfalls beachtlichen Bestände der Allgemeinen Deutschen Biographie (ADB) rund 26.500 Einträge zu Personen bis einschließlich des 19. Jahrhunderts, sowie die Neue Deutsche Biographie (NDB) derzeit knapp 22.000 Einträge.

Für erste quantitative Betrachtungen werden daher bewusst zunächst die Artikel der Wikipedia und die strukturierten Daten aus Wikidata verwendet, um eine breite Datenbasis zu schaffen. Durch die angestrebte Kombinierbarkeit und Selektierbarkeit von Quellen wird die Implementierung später auch Möglichkeiten bieten, Analysen auf historisch fundierte Quellen einzuschränken oder diese z. B. auch mit den Ergebnissen aus der Wikipedia zu vergleichen.

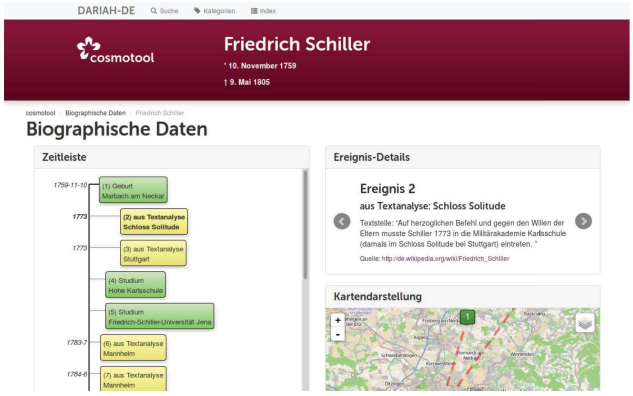
Qualitative Unterstützung der Forschung

Ein erster entwickelter Prototyp umfasst neben rund 1,8 Millionen aus Wikidata abgeleiteten, biographisch relevanten Daten auch Ergebnisse der quantitativen Analyse biographischer Texte aus der Wikipedia. Durch die Zusammenführung von Ereignissen aus unterschiedlichen und idealerweise auch mehrsprachigen Quellen werden die biographischen Profile schrittweise erweitert und verfeinert.

Kontextspezifische Kombination von Daten

Durch die Kombination unterschiedlicher Quellen kann aber nicht nur eine größere Menge an Ereignissen erkannt werden, auch die Qualität der abgeleiteten Profile kann gesteigert werden. Angaben zu Zeitpunkten, Orten und interagierenden Personen werden in unstrukturierten Texten durch die Anwendung computerlinguistischer Verfahren zwar erkannt, entsprechende Algorithmen können aber Bezeichnungen und Zusammenhänge oft nicht zweifelsfrei auflösen. Wenn nun die Analyse von Texten unterschiedlicher Sprachen und Herkunft Korrelationen erkennt, die einer gegenseitigen Plausibilitätsprüfung standhalten, so kann für entsprechende Ereignisse mit einer höheren Wahrscheinlichkeit angenommen werden, dass diese auch richtig erkannt wurden.

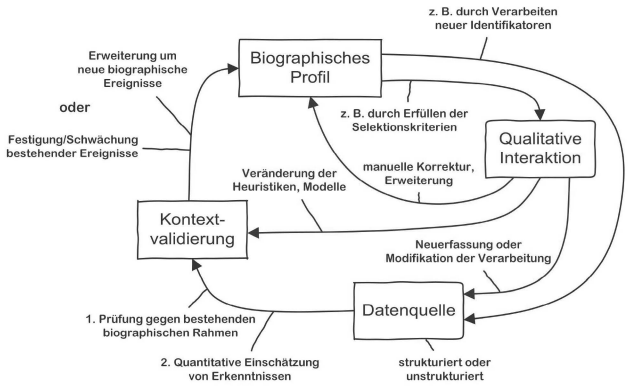
Ein einfaches Beispiel: Die Abbildung zeigt einen Überblick über erkannte Ort / Zeit-Korrelationen im Lebenslauf Friedrich Schillers. Schiller wurde nach den Angaben in Wikidata 1759 geboren. Gegen diese Information können nun die Ergebnisse von Volltextanalysen so geprüft werden, dass algorithmisch erkannte Ereignisse für das Leben Schillers in den Jahren 1710 oder 1880 als unplausibel erkannt werden. Die Farbgebung der Ereignisse in der Zeitleiste deutet die Sicherheit der Einträge an: grün steht hierbei für gesicherte Erkenntnisse, gelbe Knoten deuten auf ein unbelegtes Ereignis aus der quanitativen Textanalyse hin. Der steigendem Abstand der Knoten von der Zeitleiste spiegelt eine steigende Unsicherheit der Ereignisse im Kontext des biographischen Rahmens wider.



Iterativer Verarbeitungsprozess

Die Umsetzung des Prototypen basiert auf einem generischen Framework für die Korrelation, Verarbeitung und Transformation von Daten, welches ursprünglich für die generische Suche von DARIAH-DE entwickelt wurde und dieser auch zu Grunde liegt. Das Framework zeichnet sich insbesondere dadurch aus, dass eine Phase der deskriptiven Datenmodellierung von der Spezifikation der Verarbeitungslogik getrennt wird (vgl. Gradl / Henrich 2014). Im Wesentlichen wird dadurch erreicht, dass geisteswissenschaftliche Experten die Forschungsdaten ihrer jeweiligen Disziplin um expliziertes Wissen zum Erstellungskontext der Daten anreichern können. Andere Forschende können auf Basis der angereicherten Datenbeschreibung nun Verarbeitungsregeln so spezifizieren, dass die erweiterten Daten in den gewünschten Verwendungskontext transformiert werden. [2]

An konkreten Beispiel der Verarbeitung biographischer Daten resultiert die Anwendung des Frameworks und des zu Grunde liegenden Konzepts in einer iterativen, kontextspezifischen Verarbeitungslogik, die in der folgenden Anwendung skizziert wird und das Zusammenspiel zwischen qualitativer Forschung und quantitativen Verfahren am Beispiel des *Cosmobilities* Prototypen verdeutlicht.



So haben Forscher die Möglichkeit an drei Stellen des Prozesses manuell einzuwirken und die quantitative Verarbeitung zu beeinflussen: Zunächst werden durch die Erfassung einer Datenquelle bzw. der Beschreibung ihrer Datenstrukturen (deskriptive Datenmodellierung) biographische Daten und Texte erfasst. Erkenntnisse, die durch eine angewendete Transformation der Daten extrahiert werden können werden ggf. in den Kontext bestehender biographischer Rahmenbedingungen gesetzt und in biographische Profile übernommen.

Auf eben diese qualitativen Einschätzung können Forscher an zwei wesentlichen Stellen einwirken: Einerseits besteht die Möglichkeit, die Einordnung biographischer Daten durch die Beschreibung von Modellen und Heuristiken zu beeinflussen. Eine vereinfachte Heuristik wird in der folgenden Abbildung dargestellt. Hier würde beispielsweise ein Versterben der Mutter zu einem Eintrag im biographischen Profil des Kindes führen, welcher den Aufenthaltsort des Kindes, insofern dieses zu diesem Zeitpunkt höchstens 16 Jahre alt war, mit einer hohen Wahrscheinlichkeit mit dem Sterbeort der Mutter korreliert. An Stelle einer solchen einfachen Heuristik könnten auch komplexere, epochenspezifische Betrachtungen, wie z. B. den Lebensalterdarstellungen von Wirag (Wirag 1994) oder Anwendungen von Lebensstufenmodellen (z. B. von Grayerz 2010) nach Anforderungen der jeweiligen Forscherperspektive stehen.

```
// Child born
getClaimsForRelatives(h, h.getChild(), true, false, 0, 0, 0.9, "Kind geboren");

// Child died
getClaimsForRelatives(h, h.getChild(), false, true, 0, 30, 0.9, "Kind verstorben");
getClaimsForRelatives(h, h.getChild(), false, true, 31, 40, 0.7, "Kind verstorben");

// Spouse died
getClaimsForRelatives(h, h.getSpouse(), false, true, 0, 0, 0.7, "Partner verstorben");

// Parents died
getClaimsForRelatives(h, h.getMother(), false, true, 0, 16, 0.9, "Mutter verstorben");
getClaimsForRelatives(h, h.getFather(), false, true, 0, 16, 0.9, "Vater verstorben");
```

Die zweite Möglichkeit der qualitativen Beeinflussung besteht in der konkreten Veränderung des biographischen Rahmens, also die manuelle Erfassung oder Korrektur wesentlicher Eckpunkte wie Geburts- und Sterbedaten der einzelnen Person oder auch seiner nächsten Verwandten. Ein weiterer Iterationszyklus folgt schließlich, wenn ein verändertes Profil die definierten Selektionskriterien einer Forscherin erfüllt und in deren Fokus rückt bzw. wenn ein nun erweitertes Profil neue Hinweise auf weitere Datenquellen beinhaltet. Solche Daten können IDs in Datenbanken sein, aber auch die Vervollständigung eines Geburtsname / Geburtsdatum-Tupels, auf dessen Basis die Suche nach weiteren biographischen Texten fortgesetzt werden kann.

Ausblick

Weitere Entwicklungsschritte sind notwendig um den beschriebenen Verarbeitungszyklus im Rahmen des Prototypen vollständig abzubilden und die Interaktion zwischen qualitativen Verfahren und der qualitativen Forschung anbieten zu können.

Parallel hierzu werden derzeit auch Möglichkeiten zur Aggregation individueller Profile untersucht, um Rückschlüsse über die Transnationalität von Personengruppen anbieten und entsprechende Internationalitätskriterien ableiten zu können.

Fußnoten

1. Für eine differenzierte historische Betrachtung des Themas verweisen wir an dieser Stelle auf das Werk von Deacon, Russel und Woolacott (2010).
2. Weitere theoretische Überlegungen finden sich in Gradl / Henrich (2014); eine Ausarbeitung, die sich mit diesem Konzept technisch weiterführend auseinandersetzt wird derzeit vorbereitet.

Bibliographie

Bamman, David / Smith, Noah A. (2014): "Unsupervised Discovery of Biographical Structure from Text", in: *Transactions of the Association for Computational Linguistics* 2: 363-376.

Blessing, André / Kuhn, Jonas (2014): "Textual Emigration Analysis (TEA)", in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation* 2089-2093.

Deacon, Desley / Russel, Penny / Woolacott, Angela (2010): *Transnational Lives*. Biographies of Global Modernity. 1700-present. Basingstoke / Hampshire: Palgrave Macmillan.

Gradl, Tobias / Henrich, Andreas (2014): "A novel approach for a reusable federation of research data within the arts and humanities", in: *Digital Humanities 2014*. Book of Abstracts, Ecole polytechnique federale de Lausanne; Lausanne: Université de Lausanne 382–384 http://dh2014.org/program/abstracts/ [letzter Zugriff 09. Oktober 2015].

Grayerz, Kaspar von (2010): *Passagen und Stationen*. Lebensstufen zwischen Mittelalter und Moderne. Göttingen: Vandenhoeck & Ruprecht.

Heilbrunn, Bernice (2011-201): "Jacob H. Schiff", in: Hoyt, Giles R. (ed.): *Immigrant Entrepreneurship*. German-American Business Biographies 1720 to the Present. Vol. 3. Washington: German Histprical Institute http://immigrantentrepreneurship.org/entry.php?rec=41 [letzter Zugriff 07. Februar 2016].

Lei, Tao / Long, Fan / Barzilay, Regina / Rinard, Martin (2013): "From Natural Language Specifications to Program Input Parsers", in: *The 51st Annual Meeting of the Association for Computational Linguistics* 1294-1303.

Wikipedia (22.11.2015): "Jacob H. Schiff" https://en.wikipedia.org/wiki/Jacob_Schiff [letzter Zugriff 07. Februar 2016].

Wikipedia (07.02.2016): "Jakob Heinrich Schiff" https://de.wikipedia.org/wiki/Jakob_Heinrich_Schiff [letzter Zugriff 07. Februar 2016].

Wirag, Klaus T. (1994): *Cursus Aetatis*. Lebensalterdarstellungen vom 16. bis zum 18. Jahrhundert. München: Univ. Diss.