I like to PROV it! Ein Data Object Provenance Tool für die Digital Humanities

Mühleder, Peter
peter.muehleder@uni-leipzig.de
Universitätsbibliothek Leipzig, Deutschland

Hoffmann, Tracy
tracy.hoffmann@uni-leipzig.de
Universitätsbibliothek Leipzig, Deutschland

Rämisch, Florian
raemisch@ub.uni-leipzig.de
Universitätsbibliothek Leipzig, Deutschland

Der Umgang mit Forschungsdaten ist inzwischen ein wichtiger Bestandteil geisteswissenschaftlicher Forschungsprojekte (nicht nur in Digital Humanities (DH) Projekten). Diese können in zahlreichen Formen und Formaten - von einfachen Tabellendokumenten, Protokollen bis hin zu komplexen Datensets und Visualisierungen - vorliegen. Die Daten unterliegen während der Forschung häufig Veränderungsprozessen: Sie müssen aufwendig bereinigt, transformiert, kombiniert, angereichert und/oder korrigiert werden. Dieses 'Preprocessing' der Daten stellt sich oft als explorativer, dynamischer und iterativer Prozess dar, der beispielsweise immer wieder unterschiedliche Algorithmen oder Mappings einsetzt. Diese Bearbeitungsschritte erzeugen damit jeweils neue Versionen eines Datenobjekts.

Um einen Überblick zu behalten, kommen oft individuelle Lösungsstrategien zum Einsatz, welche in vielen Fällen für andere Personen schwer nachzuvollziehen sind. Die Nachvollziehbarkeit ist aber entscheidend dafür, ob die Daten später von anderen Forschenden nachgenutzt werden können. Um dieses Ziel zu erreichen sollte die Provenance von Daten (Informationen zu Herkunft, Verarbeitungsprozesse, etc.) während des Datenlebenszyklus mit erfasst werden. Sie hilft dabei, die Herkunft der Daten und deren Bearbeitung und auch eventuelle Fehler bis zum aktuellen Zustand zu verstehen. Mit anderen Worten, Provenance kann auch eine kritische Perspektive auf Daten ermöglichen (D'Ignazio and Klein 2016).

Die Provenance-Erfassung von Daten ist in verschiedenen (meist naturwissenschaftlichen und sozialwissenschaftlichen) Disziplinen bereits etabliert und Teil des Forschungsprozesses (vgl. Freire et al. 2008, Oliveira et al. 2018). Für die Nutzung im Kontext geisteswissenschaftlicher Forschung zeigt sich jedoch, dass die bereits bestehenden Lösungen oft überkomplex sind, da sie primär die Reproduzierbarkeit der Daten sicherstellen sollen (vgl. Pasquier, T. et al. 2017). Im Fall der Digital Humanities steht jedoch das Datenobjekt oft im Vordergrund: "Recording provenance information in the digital humanities [...] has to concentrate on the systematic recording of input and output data within the workflows" (Küster et al. 2011:321). Dabei ist es nicht zwingend notwendig den gesamten technischen Prozess (mit dem Ziel der Reproduktion) zu erfassen, sondern eine nachvollziehbare Erklärung der unterschiedlichen Iterationen des Datenobjekts zu liefern. In diesem Sinne betrachten wir Provenance in den DH in erster Linie als Dokumentationsaufgabe. Diese soll dabei eine Ergänzung zu oder einen ersten Schritt in Richtung komplexerer Workflow Management und Provenance Tracking Systeme darstellen.

Dieses Poster beschäftigt sich mit der Frage, wie eine derartige Dokumentation durch Forschungsdaten-Provenance in einem Digital Humanities aussehen kann und stellt ein Tool vor, welches im diggr Projekt entwickelt und eingesetzt wird. Damit werden exemplarisch mögliche Antworten auf zwei wesentliche Aspekte der Forschungsdaten-Provenance in Digital Humanities Projekten formuliert:

- Welche Informationen werden/sollten erfasst werden?
- Wie können die Informationen einfach erfasst und (menschenlesbar) wieder ausgegeben werden?

Provit (Rämisch & Mühleder 2018) wurde als ein Tool entwickelt, das einzelnen Forschenden und kleinen Gruppen helfen soll, in datengetriebenen, experimentellen Projekten den Überblick über Datentransformationsschritte zu behalten. Das Ziel von provit ist es, ein einfach zu bedienendes Tool sowohl für die manuelle als auch für die (semi-)automatische Erfassung von Provenance zur Verfügung zu stellen. Diese Provenance-Informationen sollen dabei langfristig (unabhängig von projektspezifischen Infrastrukturen) verfügbar sein. Es erstellt dafür für das jeweilige Datenobjekt (Datei) einen auf das PROV-O Vokabular (Labo et al. 2013) basierenden RDF Graphen, der Akteure, Aktivitäten und ihren Einfluss auf ein spezifisches Datenobjekten retrospektiv beschreibt. Zusätzlich werden weitere Metadaten wie Zeitpunkt der Bearbeitung und Ursprungsort der Daten erfasst. Eine Browser Applikation steht für einen einfachen und übersichtlichen Zugang zu den Provenance-Informationen bereit, welche diese aus einer Datei und der dazugehörigen Quelldatei als dynamische Timeline und Netzwerk dargestellt. Diese Darstellung dient dazu einen Überblick über die Zusammenhänge und Details der Bearbeitungsschritte der Datei zu erhalten (siehe Abb. 1) und unterstützt das Verständnis der Provenance-Daten auch für technisch weniger versierte Forschende.

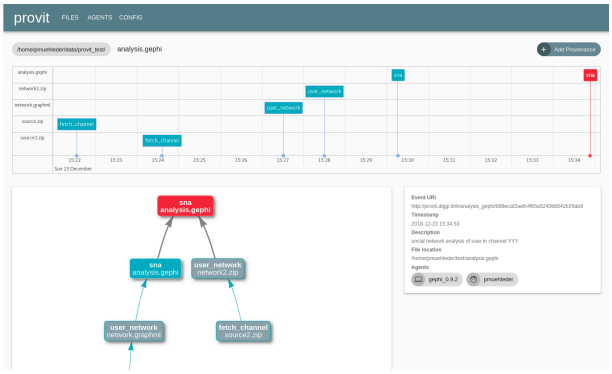

Abbildung 1. Visualisierung der Provenance-Informationen eines Datenobjektes (Provit Version 1.0 Prototyp)

Provit ermöglicht es somit auf eine standardisierte Weise, Provenance zu Forschungsdatenobjekten (unabhängig von Dateiformat und Bearbeitungsform) zu erfassen, ohne dabei auf eine komplexe technische Infrastruktur angewiesen zu sein. Die Verwendung des PROV-O Vokabulars stellt dabei die Interoperabilität sicher. Mit Provenance versehene Forschungsdaten können so dem geforderten Anspruch der Nachvollziehbarkeit Rechnung tragen und bei dem Verständnis und der Kritik von Forschungsdaten unterstützen.

Bibliographie

D'Ignazio, Catherine / Klein, Lauren F. (2016): *"Feminist Data Visualization"*, in: Workshop on Visualization for the Digital Humanities (VIS4DH), Baltimore. IEEE.

Freire, Jualina / Koop, D. / Santos, E. / Silva, C. T. (2008): "*Provenance for Computational Tasks: A Survey*", in: Computing in Science & Engineering May/June 2008, 11-21, DOI 10.1109/MCSE.2008.79.

Lebo, Timothy / Sahoo, Satya / McGuinness, Deborah (2013): *PROV-O: The PROV Ontology*. W3C Recommendation. https://www.w3.org/TR/prov-o/ [letzter Zugriff 12. Oktober 2018].

Rämisch, Florian / Mühleder, Peter (2018): *Provit* (Version v0.2.3). Zenodo, DOI http://doi.org/10.5281/zenodo.2268521

Oliveira, Wellington / De Oliveira, Daniel / Braganholo, Vanessa (2018): "*Provenance Analytics for Workflow-Based Computational Experiments: A Survey*", in: ACM Computing Surveys 51/3, 53:1-53:25, DOI 10.1145/3184900.

Pasquier, T. / Lau, M. K. / Trisovic, A. / Boose, E. R. / Couturier, B. / Crosas, M. / Seltzer, M. (2017): "*If these data could talk*", in: Scientific Data 4, DOI 10.1038/sdata.2017.114.