Richtlinie zum „Software-Sprint“

Remove NA – Knowledge Graph der LGBTIQ-Geschichte

Schlussbericht

Zuwendungsempfängerin:

Katharina Brunner

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS22S15 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

Was war Deine Motivation? Welches Problem wolltest Du mit Deinem Projekt lösen? Wie war die geplante Vorgehensweise zur Problemlösung (auch Angabe der wichtigsten Meilensteine)?

Remove NA erhöht die Sichtbarkeit von Daten zu queerer Geschichte, indem es analoge, queere Geschichte in verlinkte offene Daten verwandelt und sie in eine offene, frei zugängliche, gemeinwohlorientierte Dateninfrastruktur eingliedert.

Queere Geschichte wurde lange Zeit nicht strukturiert erfasst. Das Forum Queeres Archiv füllt seit 20 Jahren diese Lücken im analogen Raum. Die englische Abkürzung NA steht für "not available" und wird verwendet, wenn Informationen fehlen. Daten zur Geschichte von nicht-heterosexuellen Menschen sind häufig NA und das, was die nigerianische Künstlerin Mimi Onuoha als "missing data" bezeichnet: “Missing impliziert sowohl einen Mangel als auch ein Sollen: etwas fehlt, sollte aber existieren.”

Um mit maschineller Hilfe fehlende Daten aufzuspüren, braucht es ein Positiv (den Knowledge Graph zur LGBTIQ*-Geschichte), um durch einen Abgleich mit Linked-Open-Data wie GND oder WikiData das Negativ (die Leerstellen) aufleuchten zu lassen. Sie können dann in das System offener digitaler Informationen eingewebt werden, um die Vergangenheit für die Zukunft zu erhalten.

Beitrag des Projektes zu den Zielen der Förderinitative „Software-Sprint“

Wer ist die Zielgruppe für Deine Lösung? Wie profitiert sie von den Ergebnissen Deines Projekts? Welche Bezüge gibt es zu den Themenfeldern und Zielen des Software Sprints?

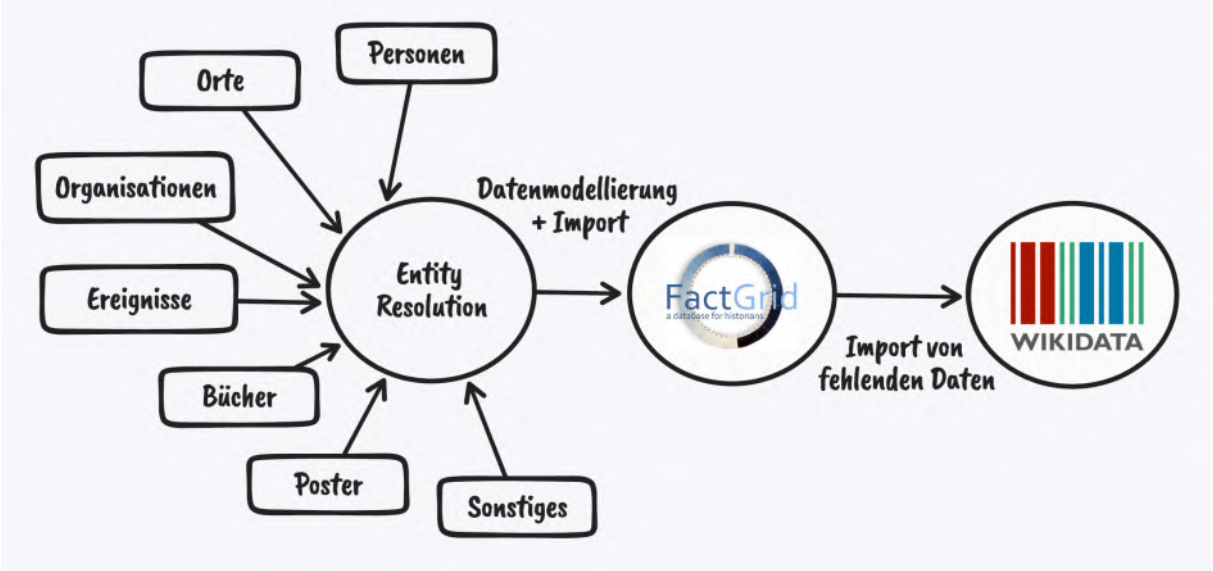
- Nutzer:innen des Archivs: Forschende und interessierte Laien, Angehörige der queeren Communities, die durch den Knowledge Graph Zusammenhänge des Bestands leichter entdecken können und möglicherweise auf hidden treasures stoßen
- Andere Community-Archive, die die technische Lösung, Dokumentation und Learnings für ihre eigenen Projekte nutzen können. Zum Beispiel ist das Queere Archiv Teil des Arbeitskreises Queer Search, in dem sich ein Dutzend ähnlicher Archive zur Digitalisierung austauscht und vernetzt
- Gemeinwohlorientierte öffentliche Dateninfrastruktur wie WikiData oder das an ein Fachpublikum gerichtete FactGrid, wo semi-automatisiert identifizierte Leerstellen ergänzt werden können, damit LGBTIQ*-Geschichte so Einzug in verschiedenste Anwendungen und Modelle finden kann

Ausführliche Darstellung der Ergebnisse

Welche konkreten Ergebnisse hast Du erzielt? Konnten alle Meilensteine erreicht werden? Welche zusätzlichen Erkenntnisse hast Du aus der Projektarbeit gewonnen, auch im Hinblick auf die Begleitung durch die Open Knowledge Foundation?

Der beste Weg um Daten und ihre Zusammenhänge und Bedeutungen für Menschen und Maschinen gleichermaßen interpretierbar zu veröffentlichen, sind verlinkte, offene Daten (Linked Open Data, kurz: LOD). Tim Berners-Lee, der Erfinder des World Wide Web, hat diese LOD als oberste Stufe eines Fünf-Sterne-Modells für offene Daten definiert.

Um mit maschineller Hilfe fehlende Daten aufzuspüren, braucht es ein Positiv (den Knowledge Graph zur LGBTIQ*-Geschichte), um durch einen Abgleich mit Linked Open Data das Negativ (die Leerstellen) aufleuchten zu lassen. Aus dem Wissen des Forums habe ich Informationen zu verschiedenen Entitäten wie Personen, Orte, Organisationen, Ereignissen, Büchern, Postern gesammelt und diese zueinander in Beziehung gesetzt. In Zahlen: Insgesamt konnte ich 7415 Entitäten aus den Datenquellen des Forums Queeres Archiv München extrahieren, die mehr als 38 000 Verbindungen haben.



Diesen internen Knowledge Graph konstruierte ich in FactGrid, das sich an ein Fachpublikum, mit akademisch-geschichtswissenschaftlichem Hintergrund richtet. Doch um sicherzustellen, dass die Daten möglichst großen, breitenwirksamen Einfluss haben, kann es nur ein Ziel in der Welt der Linked Open Data geben: Wikidata. Denn Wikidata ist die größte offene Wissensdatenbank mit aktuell 99 Millionen Objekten und das Scharnier schlechthin für externe Datenquellen. Zu jedem Objekt sind meistens nicht nur die Informationen hinterlegt, sondern auch Identifikatoren in andere Datenbanken. Jeder einzelne dieser Identifier ist ein Schlüssel in einen anderen Datensatz. Durch ein extra entwickeltes Vergleichstool, das automatisiert übereinstimmende und unterschiedliche Daten aufzeigt, konnte ich mehr als 11 200 Lücken in Wikidata schließen.

Alle Personen, die sich in diesem Datensatz zur queeren Geschichte befinden, existierten nicht für sich allein. Sie waren und sind umgeben von Menschen, die sie lieben, mit denen sie zusammenarbeiten, von denen sie verfolgt wurden, mit Menschen, von denen sie beeinflusst und solchen, die von ihnen inspiriert wurden. Kurzum: ihre Weggefährt*innen.

Nimmt man nur die Daten aus *Remove NA*, wird man diesem Beziehungsgeflecht nicht gerecht. Erst im Zusammenspiel mit anderen Datenquellen kann man sich der Realität halbwegs annähern. Sinnvoll ist es, Wikidata und Wikipedia zu verbinden. DBpedia ist ein Knowledge Graph, der von Wikipedia abgeleitet und mehrmals im Jahr aktualisiert wird.

Ein Beispiel: Bei Magnus Hirschfeld weiß man von zwei Partnern, Karl Giese und Li Shiu Tong. Beide werden im Fließtext in Wikipedia erwähnt, doch in den strukturierten Daten ist das anders. Der eine, Karl Giese, war nur in Wikipedia/DBpedia hinterlegt, der andere, Li Shiu Tong, bis zum Juli 2022 nur in Wikidata. Bis die Lücke geschlossen wurde, musste man beide Quellen abfragen.

Um die Abfrage zu persönlichen Beziehungen zwischen Menschen zu erleichtern, habe ich eine Shiny-Webanwendung gebaut, die über Federated Queries mehrere Datenbanken abfragt.

Weitere Perspektiven und Abfragen auf die Daten:

- Chronik queerer Geschichte
- München als Zentrum bayerischer LGBTIQ*-Geschichte
- Nationalsozialismus
- Beziehungsgeflechte

Die dabei gegangenen Schritten entsprechen konzeptionell den im Antrag avisierten Meilensteinen, wenn auch zum Teil mit angepassten technischen Umsetzungen und Methoden:

- Erstellen der ersten Iteration des LGBTIQ*-Knowledge-Graphs (LGBTIQ*-KG) auf Basis möglichst vieler, gesäuberter Bestandsdaten des Archivs. Kernentitäten: Person, Organisation, Ort, Zeit. Record Linkage insbesondere zu Wikidata
- Speicherung und Veröffentlichung bei FactGrid
- Lücken finden zwischen FactGrid und Wikidata, also größten freizugänglichen Knowledge Graph der der Welt durch einen strukturierten Abgleich und Massenimporten, beim Auffinden von Lücken

- Visualisierungen und Beispielabfragen bauen, um Perspektiven auf den Graph zu bekommen, siehe queerdata.forummuenchen.org

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Welcher Nutzen ergibt sich für die Zielgruppe aus den Ergebnissen Deines Projekts? Welche weitergehenden Effekte ergeben sich aus der Open-Source-Stellung der Ergebnisse? Gibt es Ideen für die Weiterentwicklung Deiner Lösung und Pläne zu deren Umsetzung?

Durch das Einweben in Wikidata stehen die Ergebnisse erstmal weiterhin für verschiedenste Anwendungen und Modelle zur Verfügung. Durch die Eingliederung in FactGrid ist ggf. auch ein Andocken an akademische Projekte leichter möglich.

Hat die Arbeit in dem Projekt Dich in Deiner persönlichen, fachlichen Weiterentwicklung unterstützt?

Ja, ich konnte mich durch Remove NA in das Themengebiet semantische Daten/Knowledge Graphs einarbeiten.

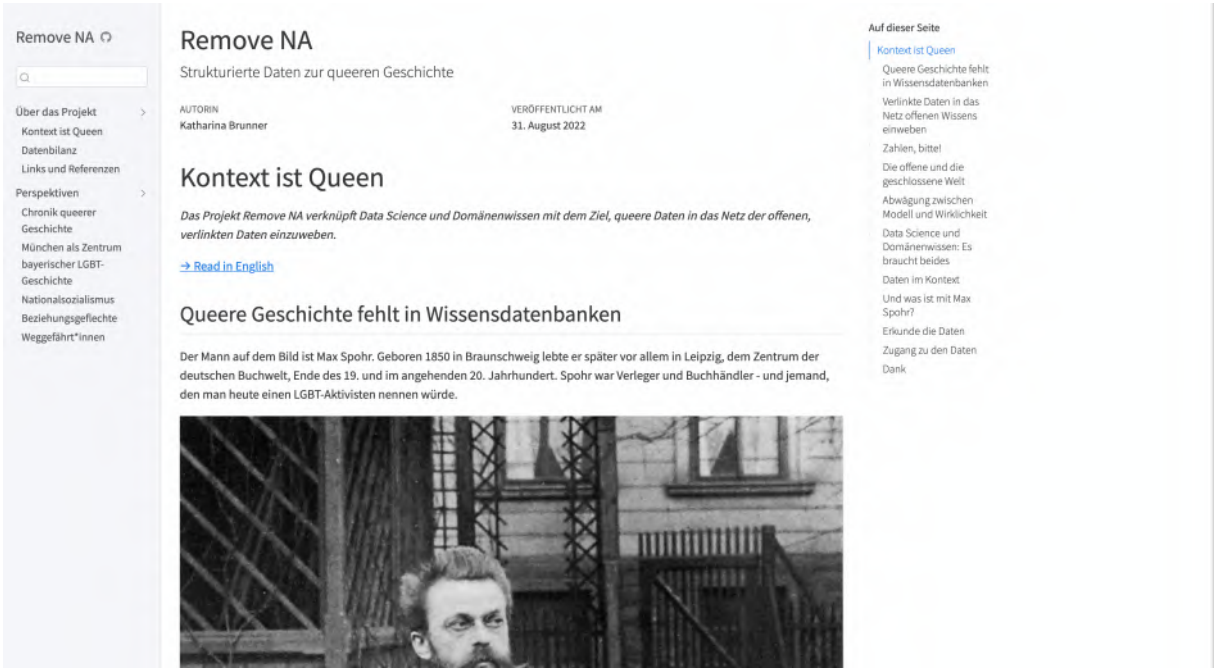
Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Gab es Arbeiten bzw. Lösungsansätze, die nicht weiter verfolgt wurden? Was waren die Hintergründe, und wie bist Du alternativ vorgegangen?

Ursprünglich hatte ich angedacht, mit Machine-Learning-Methoden die Struktur des Netzwerks genauer zu untersuchen. Dafür bietet sich etwa das Berechnen von Graph Embeddings an, um etwa Cluster zu finden. Machine-Learning-Methoden kann sich als recht aufwendig erweisen ohne dass die Ergebnisse wirklich weiterführen oder den Zusatzaufwand wert sind. Nach einer ersten Vorstudie habe ich ein ungünstiges Aufwand-Ertrag-Verhältnis erwartet und stattdessen auf klassische Netzwerkanalysen gesetzt.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Wo können sich Interessenten detailliert über Deine Projektergebnisse informieren (z.B. Webseite, GitHub, Veröffentlichungen)?



- queerdata.forummuenchen.org: Webseite mit Essay, verschiedenen Perspektiven auf die Daten und weiterführenden Informationen

- Github-Repository

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Gab es im Projektverlauf Ereignisse, die eine Anpassung der Planung erforderlich machten – z.B. Mehr- oder Minderaufwand bei der Bearbeitung von Teilaufgaben?

Nein

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Gab es Entwicklungen anderer Personen oder Institutionen, die Einfluss auf Deine Arbeiten und die Zielsetzung hatten? Wenn ja, worin bestand dieser und wie bist Du damit umgegangen?

Ja, FactGrid, eine Datenbank für semantische Daten historischer Projekte und in Person Olaf Simon, war elementarer Teil für den Erfolg des Projektes. FactGrid ist an ein Fachpublikum mit akademisch-geschichtswissenschaftlichem Hintergrund gerichtet. Der interne Knowledge Graph von Remove NA ist dort Teil der gemeinwohlorientierten und öffentlichen Dateninfrastruktur.