Richtlinie zum „Software-Sprint“

Datenland – Regionalstatistik verstehen und nutzen

Schlussbericht

Zuwendungsempfänger:
Simon Jockers und Simon Wörpel GbR

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen **01IS17S30** gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

Von den statistischen Landesämtern werden kontinuierlich große Datenmengen auf lokaler Ebene erfasst. Das reicht von Zahlen zu Arbeitslosigkeit und Sozialleistungen bis zur Tierhaltung in der Landwirtschaft. Doch politisch engagierten Bürger*innen oder Lokaljournalist*innen ist es kaum möglich, diese Daten sinnvoll einzusetzen.

Denn obwohl die Daten der statistischen Landesämter theoretisch der Öffentlichkeit zur Verfügung stehen, sind sie nur schwer zugänglich: Die offiziellen Statistikportale bieten zwar Daten zum Download an, aber die Bedienung dieser Portale und die weitere Analyse stellen Menschen ohne Fachkenntnisse vor unüberwindbare Hindernisse.

Die beiden Projektteilnehmer Simon Jockers und Simon Wörpel kennen diese Probleme auch selbst aus ihrer alltäglichen Arbeit als Datenjournalisten. Daher war die Motivation für dieses Projekt auch teilweise eine persönliche: Die eigenen Arbeitsabläufe zu verbessern, mit dem Wissen, dass viele Kolleg*innen die gleichen Probleme und Schwierigkeiten bei der Verarbeitung und Analyse dieser Datensätze erfahren.

Diese Probleme soll das Projekt “Datenguide” (gefördert unter dem Arbeitstitel “Datenland”) lösen. Es möchte Bürger*innen in die Lage versetzen, die Regionalstatistik zu verstehen und sinnvoll zu nutzen. Dazu sollte eine einfach zu bedienende Website entwickelt werden, die statistische Regionaldaten klar und prägnant darstellt, vergleichbar macht, in Kontext setzt und erklärt – beispielsweise durch interaktive Grafiken und kurze Erklärtexte. Die Website soll die Daten nach Themen strukturieren (z.B. Umwelt oder Soziales) und regional (Städte, Landkreise und Gemeinden).

Die geplanten wichtigsten Meilensteine waren daher zunächst der Aufbau einer technischen Lösung, die automatisiert Datensätze aus dem Statistikportal der Länder (www.regionalstatistik.de) importieren und verarbeiten kann. Daraus sollte eine eigene Datenbank über die einzelnen Regionen und Daten-Merkmale entstehen, die dann als Grundlage für den dritten Meilenstein, eine einfach zu bedienende Website dienen sollte. Diese Webseite sollte die Daten verständlich aufbereiten und visualisieren. In einem vierten

Schritt sollte mit den Daten inhaltlich gearbeitet werden: Einzelne Kennzahlen werden kuratiert auf der Webseite dargestellt, mit Erklärungen versehen und zur besseren Interpretierbarkeit in einen thematischen Kontext gesetzt werden.

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

Das Projekt “Datenguide” (gefördert unter dem Arbeitstitel “Datenland”) trägt zu den Zielen zweier Schwerpunkte der Förderinitiative bei: “Data Literacy” sowie “Civic Tech”. Das Projekt ermöglicht für den Bereich “Civic Tech” Bürger*innen einen deutlich vereinfachten Zugang zu Daten, die zwar grundsätzlich der Öffentlichkeit zur Verfügung stehen aber aufgrund der Art ihrer Veröffentlichung kaum von BürgerInnen verwendet werden können. Im Sinne der “Data Literacy” kann es Berührungsängste vor amtlichen Statistiken nehmen und die Hemmschwelle, diese zu benutzen, deutlich senken. Auch wird der Arbeitsaufwand für die eigene Datenrecherche und -aufbereitung für die Nutzer*innen erheblich gesenkt.

Ausführliche Darstellung der Ergebnisse

Aus diesem Projekt sind drei konkrete Ergebnisse in Form von technischen Lösungen entstanden, die unabhängig voneinander benutzt werden können und zusammen das Projekt “Datenguide” ergeben: Erstens eine Softwarelösung für das Auslesen und Aufbereiten der Datensätze aus den GENESIS-Portalen, zweitens eine Schnittstelle für die Bereitstellung der Daten zur Weiternutzung (API) und drittens ein Prototyp einer Benutzeroberfläche für die Darstellung und Kontextualisierung der Daten in Form einer Webseite (www.datengui.de), die die entwickelte API als Grundlage nutzt. Der Open-Source-Code der einzelnen Komponenten ist verfügbar unter https://github.com/datenguide

Die ersten beiden Ergebnisse, die Softwarelösung zum Herunterladen und Aufbereiten der Datensätze aus den GENESIS-Portalen sowie die Bereitstellung dieser als maschinenlesbare Schnittstelle (API) sind vollständig implementiert. Diese beiden Software-Tools können auch alleinstehend oder in anderen Projekten verwendet werden. Zusammen müssen sie verstanden werden als eine Art “Daten-Pipeline”, also eine definierte Abfolge von automatisierten Arbeitsschritten, die den Prozess des Herunterladens der Datensätze über die Aufbereitung und Umformung hin zur Bereitstellung über die API abbilden. Diese API ist seperat unter https://api.datengui.de veröffentlicht und kann frei genutzt werden.

Der dritte Meilenstein, die Website, ist technisch umgesetzt, jedoch noch nicht veröffentlicht, da die redaktionelle Bearbeitung der Datensätze noch nicht abgeschlossen ist. Voraussichtlich Anfang Mai wird die aktuelle Website des Projekts, die derzeit eine Informations-Seite über das Projekt darstellt, durch die dann fertige und mit redaktionellen Inhalten gefüllte Website ersetzt.

Somit wurden drei der vier Meilensteine erreicht, wobei beim dritten Meilenstein eine umfangreichere Ausgestaltung der Webseite geplant war. Der vierte Meilenstein, die redaktionelle Aufbereitung ausgewählter Daten auf der Webseite mit Erklärtexten, konnte im Förderzeitraum leider nicht vollständig umgesetzt werden. Die Projektteilnehmer arbeiten

allerdings über die Projektlaufzeit hinaus an der Plattform weiter und widmen sich nun vor allem den geplanten Vorhaben im vierten Meilenstein widmen.

Im Rahmen der Projektarbeiten wurden wichtige fachliche Kenntnisse erlernt. So besteht nun bei den Projektteilnehmern ein umfassendes Wissen über amtliche Statistiken in Deutschland, wie diese funktionieren und auf den verschiedenen Portalen veröffentlicht werden. Im Zuge der Programmierung der einzelnen Bestandteilen wurden zudem neue Technologien erlernt und angewendet. So werden die tabellarischen Datensätze in eine Baum-Datenstruktur umgewandelt, um in der Schnittstelle als GraphQL-Implementierung bereitgestellt zu werden. Diese Technologie war für beide Teilnehmer ein neues Feld. Auch bei der Art und Weise, wie die Webseite generiert wird, wurden erfolgreich neue Techniken eingesetzt. Die Webseite wird nun nach neuesten Standards statisch gerendert (mittels eines sogenannten “JAM-Stacks”) anstelle eines herkömmlichen dynamischen Content Management Systems. Dies ermöglicht eine ressourcensparende Veröffentlichung umfangreicher Websites.

Ebenfalls neue Erkenntnisse erlangten die Teilnehmer durch die wahrgenommenen Coachings im Bereich User Experience Design, die über die Open Knowledge Foundation angeboten wurden. Diese fachliche Weiterbildung kann nun auch im Arbeitsalltag angewendet werden.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Das Angebot von Datenguide richtet sich an zivilgesellschaftliche Initiativen und Journalist*innen. Datenguide zielt darauf ab, die Daten der Regionalstatistik einfacher nutzbar zu machen – sowohl durch eine API als auch durch eine vereinfachte grafische Darstellung auf einer Website. Im Gegensatz zu existierenden Datenportalen des statischen Bundesamts und der Landesämter (z.B. GENESIS), liegt der Fokus auf der Kontextualisierung und redaktionellen Einordnung der statistischen Inhalte und einer einfachen Konsumierbarkeit für Bürger*innen ohne besondere technische oder statistische Vorkenntnisse.

Im Gegensatz zu existierenden redaktionellen Aufbereitungen der Statistiken (z.B. dem jährlich als Buch erscheinenden Datenreport der BPB) ist der Datenguide kein statisches Dokument sondern eine interaktive Anwendung, in der Nutzer*innen auf personalisierte Informationen zugreifen können, beispielsweise auf interaktive Grafiken zu ihrer Stadt oder ihrer Region. Die Veröffentlichung des Quelltexts unter einer offenen Lizenz ermöglicht es anderen Entwickler*nnen, auf dieser Arbeit aufzubauen. Die Zerlegung in einzelne Komponenten (Daten-Pipeline, API, Website) soll es einfacher für Dritte machen, Bestandteile des Projekts in anderen Kontexten wiederzuverwenden.

Der kurzfristige Fokus der Weiterentwicklung liegt auf der redaktionellen Aufarbeitung der Daten für die Datengui.de-Website. Mittelfristig soll geprüft werden, wie die Daten automatisiert in andere Dienste exportiert werden können, beispielsweise um sie einfacher in webbasierten Visualisierungs-Werkzeugen wie “Datawrapper” nutzbar zu machen.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Zunächst war für den ersten Meilenstein geplant, eine bestehende technische Lösung für den Import und die Aufbereitung der bereitgestellten Daten zu nutzen, das Software-Tool “regenesis” (https://github.com/pudo/regenesis). Doch es stellte sich heraus, dass diese Software mittlerweile sehr veraltet und daher schwierig bis kaum mehr zu benutzen war. Es wurde abgewogen, einen gewissen Arbeitsaufwand in die Weiterentwicklung dieses Tools zu stecken, was aber schlussendlich aus ressourcengründen nicht weiter verfolgt wurde.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Das Projekt wird auf verschiedenen Kanälen für verschiedene Nutzergruppen präsentiert und zur Verfügung gestellt. So wird die Schnittstelle eigenständig unter https://api.datengui.de veröffentlicht und richtet sich an andere Software-Entwickler*innen, die basierend auf dieser API eigene Webseiten und andere Anwendungen bauen können. Des weiteren wird zu Anfang Mai die Webseite www.datengui.de mit ersten Inhalten und Datenvisualisierungen online gestellt.

Zudem werden verschiedene öffentliche Auftritte genutzt, um das Projekt bei der Zielgruppe bekannt zu machen. So stellt Simon Jockers den Datenguide auf der Medien-Fachkonferenz “re:publica” vor (https://18.re-publica.com/en/session/statistik-alle-wir-hacken-statistische-bundesamt).

Außerdem ist bereits ein Workshop bei der “Netzwerk Recherche Jahreskonferenz 2018” geplant, einer der größten und bekanntesten Konferenzen für Recherche-Journalist*innen im deutschsprachigen Raum.

Für interessierte Software-Entwickler*innen ist der komplette Open-Source-Code für die verschiedenen Komponenten des Projekts unter https://github.com/datenguide unter einer freien Lizenz veröffentlicht.

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Zu Beginn des Projektes konnten nicht die geplanten Mittel abgerufen werden, da von den Projektteilnehmern weniger Arbeitszeit aufgewendet werden konnte als zunächst geplant. Das wurde im weiteren Verlauf des Förderzeitraums wieder ausgeglichen. Deshalb wurde die zu Beginn veranschlagte Kostenplanung weitestgehend eingehalten.

Die Arbeiten am ersten Meilenstein, der technischen Lösung für Import und Aufbereitung der Datensätze aus den statistischen Portalen war aufwendiger als zunächst angenommen. Auch die Bereitstellung der eigenen Datenbank über eine maschinenlesbare Schnittstelle

(“API”) war umfangreicher. Dafür entwickelte sich diese API zu einem eigenen Zweig des Projektes, der nun als eigenständiges Ergebnis veröffentlicht wird.

Aufgrund der umfangreicheren Arbeiten an den ersten beiden Meilensteinen konnte weniger Arbeitseinsatz auf die Umsetzung der Darstellungen und Visualisierungen auf der Webseite aufgebracht werden. Die redaktionelle Aufbereitung der Daten soll über die Projektlaufzeit hinaus weiter umgesetzt werden.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Ein Vorbild für Datenguide war der Datenreport der Bundeszentrale für politische Bildung, der alle zwei bis drei Jahre vom Statistischen Bundesamt (Destatis), dem Wissenschaftszentrum Berlin (WZB) und dem Sozio-oekonomischen Panel (SOEP) des Deutschen Instituts für Wirtschaftsforschung (DIW Berlin) erarbeitet wird. Datenguide konzentriert sich im Vergleich dazu allerdings auf den Anteil der Daten aus dem gemeinsamen Statistikportal der Länder (“Regionalstatistik”) und versucht diese in einem interaktiven und regelmäßiger aktualisierten Format darzustellen. Außerdem bietet es den Zugang zu den Daten über eine API, sodass auch andere Projekte darauf aufbauen können.