DATENSCHULE

DATEN FINDEN & BEKOMMEN

OPEN KNOWLEDGE FOUNDATION DEUTSCHLAND

Mehr zu unseren Projekten & Workshops:
datenschule.de
E-Mail: info@datenschule.de |
Telefon: 030-57703666-2

DATEN FINDEN & BEKOMMEN

INHALTSVERZEICHNIS

Zivilgesellschaftliche Kampagnen und journalistische Recherchen beginnen meist mit einer Frage oder einer Hypothese. Um diese zu testen oder zu beweisen, fehlt es häufig an ausreichend Quellen und Informationen. Der richtige Datensatz kann daher ein entscheidender Faktor in einem Daten-getriebenen Projekt sein. Dieses Lehrmaterial gibt einen Überblick über hilfreiche Tipps, Tools und Datenquellen, um Daten zu finden und zu bekommen.

1. DATEN-PORTALE

Heutzutage stellen viele Regierungen und politische Institutionen Datensätze in Daten-Portalen zur Verfügung. Und immer häufiger werden sie auch aktuell gehalten und gepflegt.

Unten finden sich einige Datenportale, die für die Recherche rund um politische Themen nützlich sein können.

<u>Datenportale:</u>

Das größte Datenportal in der Deutschland:
- Govdata: https://www.govdata.de/

Karte mit Open-Data Portalen in Deutschland:
- Open Data Atlas: http://bit.ly/2sy8Ngn

Das Europäische Pendant:
- European Data Portal: https://europeandataportal.eu/

Informationen über Unternehmen und wirtschaftlich Berechtigte:
- OpenCorporates: https://opencorporates.com/
- OpenOwnership: http://openownership.org/

1. DATEN-PORTALE

Datenportale (continued):

Die Firmeninformationen der Panama Papers sind ebenfalls öffentlich verfügbar:

- Offshoreleaks: https://offshoreleaks.icij.org/

Sanktionslisten verschiedener Staaten:

- Open Sanction: www.opensanctions.org

Eine große Sammlung an Recherche-Dokumenten & Leaks hält OCCRP bereit:

- Investigate Dashboard: https://investigativedashboard.org/

Weitere interessante Datenbanken in der EU:

- Eurostat: http://ec.europa.eu/eurostat/de
- The Cohesion Fund: https://cohesiondata.ec.europa.eu/
- European Farm Subsidies: http://farmsubsidy.openspending.org

Datenbanken internationaler Institutionen (UN, OECD):

- World Bank Data: http://data.worldbank.org/
- The Data Hub: https://datahub.io/
- Data from the UN: http://data.un.org/

2. INFORMATIONSFREIHEIT

Durch das Informationsfreiheitsgesetz (IFG) werden die Grundrechte zur freien Einsicht in Dokumente von öffentlichen Einrichtungen geregelt. Jeder Bürger in Deutschland hat das Recht Informationen von öffentlichen Behörden anzufragen. Da das IFG Ländersache ist, können Anfragen nur in den Bundesländer gestellt werden, die ein IFG gesetzlich verankert haben. In dem Online-Portal FragDenStaat der Open Knowledge Foundation Deutschland e.V., können IFG-Anfragen einfach und unkompliziert gestellt werden: https://fragdenstaat.de/. Das Portal bietet auch weitere Informationen rund um das Informationsfreiheitsgesetz.

<u>Weitere hilfreiche Seiten:</u>

Wenn auch nicht ganz vollständig, Alaveteli bietet einen Überblick über IFG-Plattformen weltweit:

- Alaveteli Deployments: http://alaveteli.org/deployments/

Auf Europäischer Ebene gibt es ebenfalls ein Portal:

- AsktheEU: https://www.asktheeu.org/

3. SUCHMASCHINEN STRATEGISCH NUTZEN

Die richtige Nutzung von Suchmaschinen kann entscheidend sein, um die richtigen Ergebnisse für ein Projekt zu erhalten. Einfache Suchanfragen liefern nicht immer die richtigen Ergebnisse: Mit ein paar Tricks können Suchanfragen allerdings verfeinert werden, sodass die Chancen, passende Resultate zu bekommen, deutlich gesteigert werden.

Ein paar hilfreiche Tipps:

- Anfragen sollten so speziell wie möglich gestellt werden
- Verwende Operatoren, um bestimmte Inhalte auszuschließen (-) oder miteinzubeziehen (+)
- Mit Hilfe von Anführungszeichen werden nur Seiten angezeigt, die 100% den Suchbegriffen entsprechen
- Indem man den File-Typen festlegt (z.B. “filetype:XLS”) werden nur Ergebnisse dieses Typen angezeigt

Weitere Quellen für Suchmaschinen Tricks:

- Huffington Post: http://bit.ly/1kQVBdQ
- TechRepublic: http://tek.io/TQHWti
- SearchEngineWatch: https://searchenginewatch.com/
- BigDataUniversity: http://bit.ly/2sTnBZS
- Das Data Journalism Handbook enthält ein Kapitel mit vielen Beispielen und wie sich Datensätze besser finden lassen.

4. WEB-APIs

Eine API (Application programming interface) ist eine spezielle Schnittstelle, an der Daten von Webseiten abgefragt werden können. Twitter, Facebook und viele Andere bieten APIs an, über die Nutzer Daten erhalten.

Beispiele:

- Über die APIs der New York Times lassen sich Artikel bis ins Jahr 1851 erhalten https://developer.nytimes.com/

- Mit der Twitter API können Tweets automatisch erstellt und gepostet werden und es lassen sich Account-Informationen über Follower und Aktivität einzelner User abfragen. https://dev.twitter.com/rest/public

- Das Organized Crime and Corruption Reporting Projects (OCCRP) bietet das Investigative Dashboard mit vielen geleakten Dokumenten und Informationen aus früheren Recherchen rund um das Themen Kriminalität und Korruption.

5. WEB-SCRAPING

Beim Web-scraping werden Informationen automatisch aus dem Netz heruntergeladen. Es ist besonders nützlich, um langwieriges Kopieren und Einfügen zu vermeiden. Webseiten, die einheitlich aufgebaut sind, lassen sich am besten scrapen. Online Medien sind dafür ein gutes Beispiel. Es gibt viele Tools und Plugins, die einem beim Web-scraping helfen können.

Google Sheets:

Ein nützliches Tool um HTML strukturiert von einer Webseite zu erhalten ist Google Sheets. Mit der Funktion "=importHTML" können Tabellen und Listen aus Webseiten auslesen und automatisch in Google Sheets importiert werden.

Google Chrome Scraper Plugin:

Ein weiteres nützliches Tool ist das Web Scraper Plugin von Google Chrome. Das Plugin ist eine Browser-Erweiterung mit dem sich einzelne Elemente einer Website anwählen und anschließend auslesen lassen. Die Daten werden als CSV. exportiert. Das untenstehende Youtube Video erklärt, wie das Plugin genutzt werden kann:
https://www.youtube.com/watch?v=oCp9IcdSpZI&t=155s

6. PDF-SCRAPING

Auch PDFs können gescraped werden. Allerdings ist das komplizierter als bei einer Webseite, da es sich bei einem PDF um eine Druckdatei handelt. Die Informationen sind hier nicht mehr durch Tags strukturiert und können daher nicht so leicht angewählt werden. Mit einigen Tools lassen sich allerdings trotzdem Informationen aus PDFs strukturiert auslesen.

Tabula:

Tabula ist ein User Interface im Webbrowser und bietet die Möglichkeit, Tabellen in PDFs zu erkennen und in ein Spreadsheet-Format zu konvertieren. Wie gut Tabellen hier ausgelesen werden können, hängt zum einen von der Qualität des PDFs ab. Zum anderen auch, wie die auszulesende Tabelle strukturiert ist. Unter folgendem Link lässt sich das Tool erhalten:
Tabula: http://tabula.technology/

Tesseract:

PDFs können auch als Bilder formatiert sein. Tools wie Tabula können dann nicht mehr Tabellen automatisch erkennen. Tesseract ist ein Comandline-Tool mit dem sich PDFs in Text umwandeln lassen. Dies kann hilfreich sein, um PDFs anschließend zu scrapen. Auch hier ist die Qualität der PDF entscheidend. Tesseract kann mittels Github installiert werden:
https://github.com/tesseract-ocr/tesseract

DATENSCHULE

Die Datenschule vermittelt gemeinnützigen Organisationen die nötigen Fähigkeiten, Daten und Technologien zu verstehen, um sie zielgerichtet für ihre gesellschaftlichen Aufgaben einzusetzen.

OPEN KNOWLEDGE FOUNDATION DEUTSCHLAND

Die Open Knowledge Foundation Deutschland ist ein gemeinnütziger Verein, der sich für offenes Wissen, offene Daten, Transparenz und Beteiligung einsetzt.

Mehr zu unseren Projekten & Workshops: datenschule.de
E-Mail: info@datenschule.de |
Telefon: 030-57703666-2