Datenrecherche, warum eigentlich mit Code?

Aggregieren

- Welche Daten sind offen verfügbar?
- Sind sie maschinenlesbar?
- Sind sie via einer API abrufbar?
- Ist das Datenwörterbuch (Data Dictionary) verständlich?
- Können die nötigen Datensätze kombiniert werden?

Falls Nein:

- Können Daten gescrapt werden? Sind die Daten in einer öffentlich zugänglichen Datenbank verfügbar?
- Können Datensätze über Zeit hergestellt werden?
- Müssen Datensätze via Crowdsourcing oder einer gezielten Umfrage aggregiert werden?

Nehmen wir das Beispiel des Schweizerischen Nationalfonds (SNF). Der SNF bietet eine Datenbank an, in der alle vergangenen und laufenden Projekte abgerufen werden können. Mit Einzelbuchstaben lassen sich die Ergebnisse abrufen. Aber es wäre nun ein Fehler, auf die Idee zu kommen, ein Programm zu schreiben, um die eigene Datenbank zu bauen. Denn wer sich auf der Website noch etwas weiter hervor klickt, entdeckt, findet diesen Link. Laden wir einmal die Grants herunter. Damit werden wir später arbeiten.

Reinigen

Simple Sachen können in Spreadsheets erledigt werden. Beispielsweise Zellen nach einem bestimmten Muster teilen, den Weissraum abschneiden, etc. Beispiel.

Auch wenn es komplizierter wird, gibt es Werkzeuge, um Datensätze zu säubern, Open Refine beispielsweise. localhost:3333.

Doch wenn man grosse Datensätze hat (sehr oft auch bei den kleineren), lohnt es sich, sie mit Hilfe des eigenen Codes zu reinigen.

Das Reinigen ist zudem der Prozess, den man bei der Arbeit unterschätzt. Oft sind auch vermeintlich saubere Daten, sehr dreckig.

Analysieren

Bevor man mit der Analyse beginnt, gilt es, sich die Fragen, die man stellen möchte und die man beantworten haben möchte, aufzulisten. Wichtig ist es, nun bei der Arbeit, diese Fragen alle zuerst zu beantworten, bevor man sich neue Fragen stellt. So verhindert man, sich in den Daten zu verlieren.

Visualisieren

Für kleinere Datensätze gibt es mittlerweile sehr, sehr viele nützliche Visualisierungswerkzeuge. Angefangen bei den Spreadsheets natürlich, aber auch andere noch einfacher anzuwendende Angebote wie Datawrapper, Plotly oder sehr mächtige Werkzeuge wie Tableau. Aber wieder stösst man mit all diesen Werkzeugen irgendwann an Grenzen. Vor allem lassen sie sich nicht in einen Workflow integrieren.

Storytelling

Nun muss man sich überlegen, wie man in welchem Medium die Geschichte am besten erzählt - mit Wörten, mit Graphiken oder gar mit interaktiven Elementen.

Publizieren und Code teilen

Den gesamten Prozess, den wir hier durch gegangen sind, lässt sich in einem Arbeitsgang mit Code erledigen. Die Daten von einer Website oder API herunterholen, die Daten scrapen, reinigen, analysieren und schliesslich zu vislualieren.