Ramon Voges, André Wendler

Vom Zettelkasten zum interaktiven Notizbuch. Das Datenlabor der Deutschen Nationalbibliothek

Man stelle sich folgendes Szenario vor: Die Mitarbeiterin einer Kulturinstitution plant eine Jubiläumsfeier. Aus diesem Grund möchte sie ein Exposé mit den Lebensläufen bekannter Persönlichkeiten zusammenstellen, die alle am gleichen Tag, in diesem Fall dem 4. August, Geburtstag haben. Um sich einen ersten Überblick zu verschaffen, ruft sie zunächst auf Wikipedia den Eintrag zum 4. August auf. Nachdem sie sich etwas auf der Seite umgesehen hat, ist ihr schnell klar: Die dort versammelten Informationen händisch zusammenzutragen, wäre zu viel Aufwand. Allein bis 1850 führt die Online-Enzyklopädie über 50 Personen auf, die am 4. August geboren worden sind. Bis zum Ende des 20. Jahrhunderts kommen etwa 450 dazu. Anstatt zwischen den einzelnen Einträgen hin- und her zu springen, alles Wissenswerte auszuwählen und dann in eine Datei zu kopieren, schreibt sie ein kurzes Skript. Nachdem sie es ausgeführt hat, befindet sich auf ihrem Rechner ein Dokument, in dem alle Lebensläufe der Geburtstagskinder vom 4. August chronologisch der Reihe nach abgelegt sind. Damit arbeitet sie weiter und macht sich an die inhaltliche Auswahl.

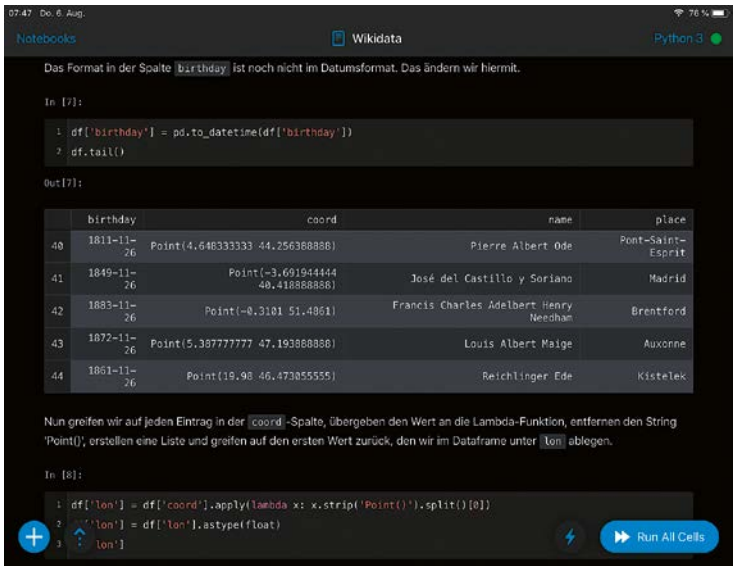
Fälle wie dieser begegnen uns in der täglichen Arbeit allenthalben. Längst nutzen Mitarbeiter*innen von Kultur- und Gedächtnisinstitutionen nicht mehr nur Bücher und Zettelkataloge, um Wissen über ihre Bestände abzurufen oder zu speichern. Stattdessen greifen sie auf Datenbanken, Online-Repositorien und Netzpublikationen zurück. Dafür allerdings müssen sie sich im Umgang mit diesen digitalen Medien ebenso gut auskennen wie mit den althergebrachten, analogen Informationsträgern. Wie man am besten an die gesuchten Informationen herankommt und gegebenenfalls mit ihnen weiterverfährt, hat sich mit dem Aufkommen elektronischer Datenverarbeitung in der Welt der Bibliotheken, Museen und Archive grundlegend geändert. Data Literacy[1], also die Fähigkeiten und Kenntnisse, die für den Umgang mit Daten notwendig sind, gehört deshalb zu den Kompetenzen, die im digitalen Zeitalter Mitarbeiter*innen von Kulturinstitutionen immer mehr im Alltag benötigen.

Programmieren im Museum?

Das Deutsche Buch- und Schriftmuseum (DBSM) dient der Deutschen Nationalbibliothek (DNB) zum einen als Schaufenster, in dem es einer breiten Öffentlichkeit vermittelt, worin die Aufgaben einer Nationalbibliothek bestehen und über welche Bestände sie verfügt. Zum anderen versteht sich das DBSM als eine Art medienhistorischer Reflexionsraum, da es den aktuellen Wandel unserer Kommunikationsformen kritisch begleitet und in eine Langzeitperspektive von 5.000 Jahren Buch- und Schriftkultur rückt. Dabei ist das Museum selbst vom medialen Wandel betroffen und sieht sich vor neue Herausforderungen gestellt. Wie, beispielsweise, soll mit Arbeiten von Typografen, mit Hausarchiven von Verlagen oder Nachlässen von Buch- und Druckhistorikern umgegangen werden, die nicht in gedruckter, analoger Form, sondern als PDFs und Word-Dokumente vorliegen? Für uns als Mitarbeiter*innen des DBSM ist es deshalb gleich doppelt geboten, dass wir uns mit dem Thema Data Literacy auseinandersetzen. Wir erzeugen nicht nur zahlreiche Metadaten zu unseren Beständen. Auch die musealen Objekte begegnen uns zunehmend in digitaler Form, also als Daten.

Als Teil einer umfassenderen Digitalstrategie haben wir deshalb das Datenlabor der DNB ins Leben gerufen. Mit ihm gibt es einen virtuellen Ort, in dem wir mit den immensen Daten der DNB spielerisch umgehen und experimentieren können, um neues Wissen über unsere Bestände zu erlangen. Darüber hinaus bietet das Datenlabor einen Rahmen, in dem wir Kolleg*innen aus anderen Kultur- und Gedächtnisinstitutionen unsere Erfahrungen im Umgang

mit den Datenschätzen der DNB weitergeben können. Das findet unter anderem in Form von Einführungen in die Programmiersprache Python statt.

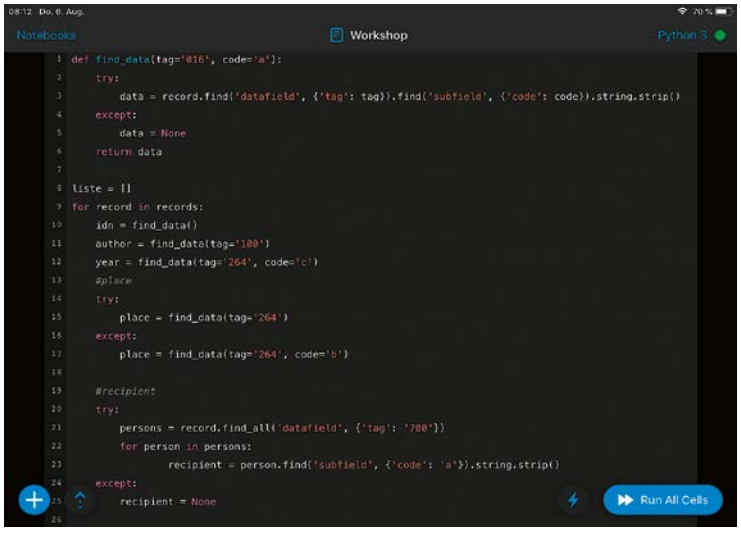


Interaktive Notizbücher

In diesen Einführungen lernen die Teilnehmer*innen die Grundlagen des Programmierens kennen. Dafür bietet sich Python besonders an, weil es einerseits eine vielseitig einsetzbare Sprache, andererseits aber auch eine leicht zu erlernende ist.[2] Die Einführungen bleiben dabei stets anwendungsbezogen und vermitteln das Wissen um Datentypen, Verzweigungen und Schleifen sowie best practices und Hilfsmittel, die den Programmieralltag leichter machen, aus dem Blickwinkel von Mitarbeiter*innen von Kultur- und Gedächtnisinstitutionen. In komplexer werdenden Schritten lernen die Teilnehmer*innen, wie sie Daten mithilfe von Python laden, verarbeiten und wieder ausgeben können. Die Anwendungsbeispiele stammen allesamt aus der alltäglichen Arbeit in Bibliotheken, Museen und anderen Kultureinrichtungen. Auf diese Weise üben die Teilnehmer*innen, das algorithmische Denken, also das Zerlegen komplexer Probleme in kleine und einfach zu lösende Einzelaufgaben, auf ihre gewohnten Tätigkeiten zu übertragen.[3] Ihren eigenen Code schreiben sie in interaktive Notizbücher, sogenannte Jupyter Notebooks.[4] Diese erlauben es den Teilnehmer*innen, Programmcode, Visualisierungen der Daten und flankierende Erläuterungen anschaulich miteinander zu verbinden. Die Dokumentation der Veranstaltungen erfolgt in etherpads, in denen die Teilnehmer*innen ihre neuen Erkenntnisse festhalten und sich mit den anderen darüber austauschen können.[5]

Zettelkasten, Web Scraping und Wikidata

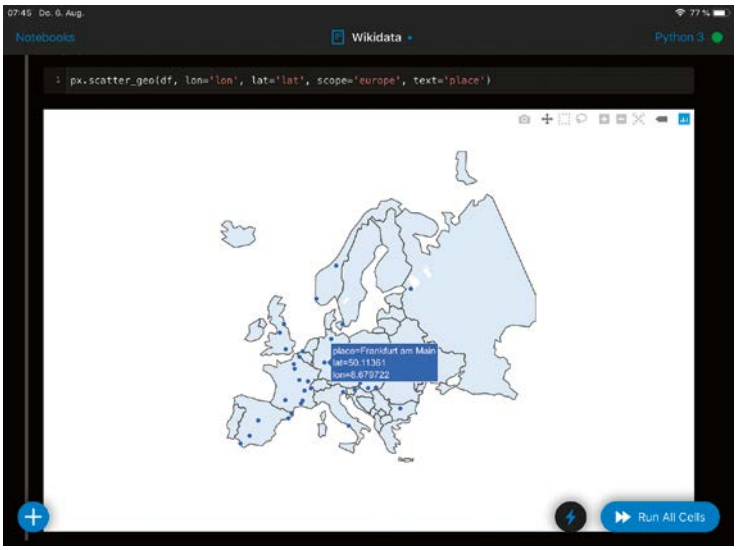
Die Einführung in das Programmieren besteht aus drei aufeinander aufbauenden Teilen. Am Anfang geht es darum, einen analogen Zettelkasten, wie er noch vor zwanzig Jahren in Bibliotheken zum Einsatz gekommen ist, digital nachzubilden. Hierbei lernen die Teilnehmer*innen einfache und komplexe Datentypen wie Zeichenketten, ganze Zahlen, Listen und dictionaries kennen. Mithilfe von Schleifen und Verzweigungen ist es ihnen am Ende dieses ersten Teils möglich, einen digitalen Zettelkasten mit Suchabfrage zu programmieren.
Im zweiten Teil setzen sich die Teilnehmer*innen mit dem Thema Web Scraping auseinander, das heißt mit dem automatischen Gewinnen von Informationen, die auf Webseiten hinterlegt sind.[6] Dafür erhalten sie einen Einblick darin, wie Internetseiten und die Auszeichnungssprache HTML funktionieren.[7] Sie lernen, mithilfe von Python die Inhalte von Seiten auszulesen, einzelne HTML-Elemente gezielt anzusteuern und deren Inhalte in eigene Datenformate zu übertragen. Darüber hinaus geht es darum, wie man programmatisch mit Fehlern umgeht und verhindert, dass das Programm abstürzt.



Auslesen von Marc21 mit Python-Code

Der dritte Teil baut auf diesen Fähigkeiten auf, wendet sich aber wieder einer neuen Informationsquelle zu, nämlich Linked Open Data.[8] Neben einer Einführung in die Abfragesprache SPARQL erfahren die Teilnehmer*innen, wie sich mit Python Daten aus Wikidata gewinnen und auf unterschiedliche Weise visualisieren lassen.[9] Am Ende der Einführung kennen die Teilnehmer*innen nicht nur die Grundlagen des Programmierens, sondern können auch mithilfe von Python Daten aus diversen Quellen laden, weiterverarbeiten und in einem adäquaten Format ausgeben.[10]

Die Corona-Pandemie hat es bislang unmöglich gemacht, die Einführungen in Python weiter in den Räumen der DNB anzubieten. Da jedoch die Nachfrage nach diesen Veranstaltungen ungebrochen ist, arbeiten Mitarbeiter*innen im DBSM daran, das Workshop-Format in ein Online-Tutorial zu übertragen, das ebenfalls auf der Grundlage von Jupyter Notebooks laufen wird.



Karten für die Visualisierung von Daten

Anmerkungen

1 Zu Data Literacy vgl. auch den Eintrag unter <https://de.wikipedia.org/wiki/Datenkompetenz>

2 Vgl. dazu beispielsweise den Eintrag bei Wikipedia unter <https://de.wikipedia.org/wiki/Python_(Programmiersprache)> sowie die offizielle Seite von Python <https://www.python.org/>.

3 Vgl. unter anderem <http://d-nb.info/1123227381, S. 89–92>..

4 Vgl. <https://jupyter.org/>.

5 Vgl. <https://de.wikipedia.org/wiki/Etherpad>.

6 Vgl. dazu die bekannte Seite <https://de.wikipedia.org/wiki/Screen_Scraping>.

7 Vgl. dazu die bekannte Seite <https://wiki.selfhtml.org/>.

8 Vgl. <https://de.wikipedia.org/wiki/Linked_Open_Data>.

9 Zu SPARQL vgl. <https://de.wikipedia.org/wiki/SPARQL>.

10 Frühere Versionen des Lernstoffs sind auf dem Github-Repositorium des DBSM zu finden: <https://github.com/buchmuseum/datenlabor_2019>.