

Hands-on-Übung zur Datenbereinigung mit



OpenRefine

Agnes Brauer
a.brauer@ub.uni-frankfurt.de

Vorbereitung | Hinweise

- Melden Sie sich für den moodle-Kurs [Praxislabor Digitale Geisteswissenschaften](#) an und schreiben Sie sich ein:

Praxislabor Digitale Geisteswissenschaften: Einführungsworkshops zu Methoden und Werkzeugen der Digital Humanities

Praxislabor Digitale Geisteswissenschaften: Einführungsworkshops zu Methoden und Werkzeugen der Digital Humanities

Zum ersten Kennenlernen von Methoden und Werkzeugen der Digital Humanities bietet die Universitätsbibliothek JCS (im Bibliothekszentrum Geisteswissenschaften) Studierenden und Mitarbeiterinnen der Goethe-Uni im kommenden Wintersemester Workshops an. In niederschweligen Einführungen werden anhand von überschaubaren, konkreten Beispielen aus der Praxis Methoden, Tools oder Themen der digitalen Geisteswissenschaften vorgestellt und geübt und so ein erster Einblick in die Möglichkeiten gegeben, wie klassische Methoden der Geisteswissenschaften mithilfe digitaler Verfahren der Textanalyse sowie der Text- und Datenaufbereitung sinnvoll ergänzt werden können.

Die Workshopreihe besteht jeweils aus inhaltlich zusammenhängenden Zweierblöcken, in denen auf eine Präsentation eine Sitzung zur Vertiefung und Übung folgt.

Die Workshops richten sich an interessierte Einsteiger; besondere Kenntnisse werden nicht vorausgesetzt. Nähere Informationen sowie die Möglichkeit zur Anmeldung finden Sie unter: <http://www.ub.uni-frankfurt.de/digitalhumanities/workshops.html>.

Praxislabor Digitale Geisteswissenschaften: Einführungsworkshops zu Methoden und Werkzeugen der Digital Humanities

Startseite / Kurse / Verschiedenes / Praxislabor Digitale Geisteswissenschaften

Allgemeines

In dieses kollaborative Dokument können Themenvorschläge für die Hands-on-Sessions eingetragen werden.

Informationen zur Anmeldung und Kurszeiten unter: <http://www.ub.uni-frankfurt.de/digitalhumanities>

Einführung in TEI / XML

Dozentin: Agnes Brauer

Der Workshop führt in die Grundlagen der Textauszeichnung mit TEI ein, einer XML-basierten und sich mittlerweile als De-facto-Standard etablierten Auszeichnungssprache speziell für die Zwecke der Geisteswissenschaften. Nach einer knappen allgemeinen Einführung werden die Teilnehmer anhand einer kleinen Übung die Praxis der Textauszeichnung mit TEI kennenlernen und sich einen ersten Überblick über die Bedeutung und die verschiedenen Module dieser Sprache verschaffen.

Link: <http://www.tei-c.org/>

Hands-on Übung zur TEI/XML-Einführung

Dozentin: Agnes Brauer

Einführungsworkshops zu Methoden und Werkzeugen der Digital Humanities

Praxislabor Digitale Geisteswissenschaften: Einführungsworkshops zu Methoden und Werkzeugen der Digital Humanities

Zum ersten Kennenlernen von Methoden und Werkzeugen der Digital Humanities bietet die Universitätsbibliothek JCS (im Bibliothekszentrum Geisteswissenschaften) Studierenden und Mitarbeiterinnen der Goethe-Uni im kommenden Wintersemester Workshops an. In niederschweligen Einführungen werden anhand von überschaubaren, konkreten Beispielen aus der Praxis Methoden, Tools oder Themen der digitalen Geisteswissenschaften vorgestellt und geübt und so ein erster Einblick in die Möglichkeiten gegeben, wie klassische Methoden der Geisteswissenschaften mithilfe digitaler Verfahren der Textanalyse sowie der Text- und Datenaufbereitung sinnvoll ergänzt werden können.

Die Workshopreihe besteht jeweils aus inhaltlich zusammenhängenden Zweierblöcken, in denen auf eine Präsentation eine Sitzung zur Vertiefung und Übung folgt.

Die Workshops richten sich an interessierte Einsteiger; besondere Kenntnisse werden nicht vorausgesetzt. Nähere Informationen sowie die Möglichkeit zur Anmeldung finden Sie unter: <http://www.ub.uni-frankfurt.de/digitalhumanities/workshops.html>

Trainerin: Agnes Brauer
Trainerin: Jakob Frohmann

Selbsteinschreibung (Teilnehmer/in)

Kein Einschreibeschlüssel notwendig

EINSCHREIBEN

Vorbereitung und Hinweise für die Hands-on-Übung

- Laden Sie bitte OpenRefine (3.1) herunter, entpacken und installieren Sie die Software auf Ihrem Computer. Sie benötigen den Browser Firefox, in dem OpenRefine läuft:
<http://openrefine.org/download.html>
- OpenRefine ist eine Java-Anwendung → es wird eine Java-Laufzeitumgebung benötigt (bitte installieren, falls auf Ihrem Gerät noch nicht vorhanden: <http://java.com/> bzw. Open Source Variante:OpenJDK
<https://github.com/ojdkbuild/ojdkbuild>
- Starten Sie die Anwendung aus dem entpackten Verzeichnis, es öffnen sich eine Kommandozeile und kurz danach der Browser mit dem geladenen Programm – sollte der Browser nicht starten, benutzen Sie bitte den Link:
<http://127.0.0.1:3333/>.
- Hilfe / Infos zum Setup: <https://librarycarpentry.org/lc-open-refine/setup.html>

Heute ...

- Zellen teilen und (wieder) vereinigen mit Hilfe von Separatoren
- Facetieren, Filtern + Clustern von Daten
- Anreichern eigener Daten aus externen Quellen (Beispiel: Wikidata)

Einfaches Beispiel zum „Aufräumen“ von Daten

Beispiel: Schlagworte zum Thema „Bestandserhaltung“

- Erstellen Sie ein OpenRefine-Projekt mit Daten der Übungsdatei „BE.txt“
- Legen Sie eine einspaltige Tabelle an, in der in jeder Zeile ein Schlagwort steht („Edit Cells“ → „Split multi-valued cells...“ → als Separator ein Leerzeichen wählen)
- Verschaffen Sie sich einen ersten Überblick über die Daten mit Hilfe der Funktion „Text facet“
- Welche Schlagworte kommen im Datensatz mehrmals vor?
- Probieren Sie auf der Grundlage der Facette die Funktion „Cluster“ aus und beseitigen Sie Tippfehler
- Vergleichen Sie die Ergebnisse beim Einsatz unterschiedlicher Clustering-Methoden und -Algorithmen

Übung 2: „Aufräumen“

- Erstellen Sie das OpenRefine-Projekt mit Daten aus dem Dokument „Muenzen.txt“
Vgl. <https://ikmk.smb.museum/object?lang=de&id=18206726>
- Explorieren Sie die Daten und betreiben Sie etwas Datenbereinigung mithilfe der Funktionen „Text-Facet“ und „Cluster“
- Transponieren Sie die Daten in mehrere Spalten: **Transpose** → **Transpose cells in rows into columns...**
- Wie viele Spalten sind sinnvoll?
- Überlegen Sie sich weitere sinnvolle „Aufräumarbeiten“ und probieren Sie sie einfach aus

Anwendungsmöglichkeiten – fortgeschrittene Funktionen

Reconcile & Match

- Vergleichen / Angleichen der eigenen Daten anhand von Datenbanken (z.B. Wikidata)
- Anreicherung von Daten (z.B. mit Identifiern oder geographischen Koordinaten)
- Verlinkung von Daten

Reconciliation is the process of matching name strings to identifiers of entities in a database like an authority file, Wikidata etc. This is useful whenever you want to merge differing name strings for the same person in your data or when you want to fetch additional data from the target database you are reconciling against.

Reconcile & Match

- Legen Sie ein neues Projekt anhand der Datei „Präsidenten.txt“ an
- Bereinigen Sie die Daten
- Splitten Sie die Daten mithilfe eines geeigneten Separators
- Führen Sie eine Reconciliation mit Wikidata durch
- Fügen Sie neue Spalten auf Basis der Reconciliation hinzu, z.B.:
 - Beruf
 - Geburtsort
 - Koordinaten des Geburtsortes
 - Geburtsdatum

Reconcile & Match

- Bereiten Sie die Datei „Praesidenten2.csv“ für den Dariah-Geobrowser auf:
<https://geobrowser.de.dariah.eu/index.html>
- Aufbau der Tabelle:

	A	B	C	D	E	F	G	H	I
1	Name	Address	Description	Longitude	Latitude	TimeStamp	TimeSpan:begin	TimeSpan:end	GettyID
2									
3									
4									
5									
6									
7									

Tipps & Tricks / Links / Literatur

Blogs

<https://histhub.ch/cat/net/blog/openrefine/> (Blog-Serie zur Arbeit an historischen Daten mit OpenRefine)

<http://blog.lobid.org/2018/08/27/openrefine.html> (einfache Anreicherung von Daten in OpenRefine mit [Personen-] Daten aus der GND via lobid.org)

Literatur

Ruben Verborgh/Max de Wilde, Using OpenRefine. The essential OpenRefine guide that takes you from data analysis and error fixing to linking your dataset to the Web (Community experience distilled), Birmingham, Mumbai 2013. [[Online-Ressource über UB FFM](#)]

Danke für Ihre Aufmerksamkeit!

Vielen Dank für die zur Verfügungstellung von Materialien und Daten an Jakob Frohmann (Universitätsbibliothek Johann Christian Senckenberg) und und Jun.-Prof. Dr. Torsten Hiltmann (Zentrum für Digitale Geschichtswissenschaft, Universität Münster).

Workshop konzipiert in Anlehnung an "[Library Carpentry: OpenRefine Lessons for Librarians.](#)" (2016)