# Automat. Ortsnormdaten - Anreicherung über regionale Eingrenzung

# Voraussetzungen

* Python und verwendete Pakete
* EAD-XML Dateien, die Ortsnamen unter //c/index/indexentry/geogname verzeichnen

# Kurzanleitung

* Die EAD-Dateien in einem Unterordner von 'data' zusammen mit der Datei 'Anreicherung.xlsx' ablegen.
* In das Feld 'B2' die bzw. mehrere Geonames-ID einer Gebietskörperschaft angeben (für Karlsruhe etwa 3214104) für die, die Kind-Orte zur Anreicherung herangezogen werden sollen. Diese Gebietskörperschaft darf aber nicht zu groß gewählt werden (z.B. BW, siehe Einschränkungen), da die Daten in diesem Prototyp über Webschnittstellen bezogen werden.
* Skript starten. Die Ergebnisse befinden sich in einer vom Ordner 'data' gespiegelten Struktur unter 'data\_enriched' bzw. in der csv-Liste und der Statistik.xlsx unter 'data\_geogname'.

# Möglichkeiten

* Verarbeitung von EAD-Dateien, um diese außerhalb des Erschließungssystems anzureichern, mit gleichzeitiger Ausgabe von csv-Listen um die Möglichkeit zu haben, die Anreicherungsinformationen in die eigene Datenbank zurückspielen zu können.
* Eingrenzung der Anreicherungsinformationen auf bestimmte geographische Bereiche, um Ortsnamen eindeutig identifizierbar zu machen. Es können auch mehrere Geonames-ID angegeben werden, um etwa bei einem Stadtarchiv neben der Stadt auch den umgebenden Landkreis zu berücksichtigen.
* für ein Archiv können mehrere 'Anreicherung.xlsx' Dateien und Ordner erstellt werden, wenn es etwa unterschiedliche räumliche Eingrenzungen gibt (also zum Beispiel ein Fotobestand aus einer Provinz in Siebenbürgen ebenfalls im Archiv vorhanden ist). Die EAD-Dateien müssen manuell in die entsprechenden Ordner kopiert werden.
* Möglichkeit nach einer ersten Analyse die Ergebnisse zu kontrollieren und nur ausgewählte Orte anzureichern, bzw. noch zusätzliche Einträge zur Anreicherung manuell aufzunehmen
* Zur Anreicherung wird die Geonames-ID herangezogen (immer) bzw. der Link auf die deutsche Wikipedia und die GND über Konkordanzen zu bestimmen versucht.
* Statistische Auswertung über alle abgelegten EAD-Dateien hinsichtlich Ortsindexbegriffen, um zu analysieren für welche Archive eine derartige Anreicherung sinnvoll ist
* Statistische Auswertung der Anreicherungsergebnisse bzw. der manuellen Kontrolle
* Am Anfang des Skripts können diverse Analyse- und Anreicherungsmechanismen ausgeschaltet werden, um das Skrpt für Analysezwecke schneller zu machen.

# Funktionsweise

* Es werden nur jene XML-Dateien verarbeitet, bei denen sich im gleichen Ordner eine Datei mit dem Namen 'Anreicherung.xlsx' befindet.
* Das Skript verarbeitet jene Inhalte die in den abgelegten XML-Dateien unter //c/index/indexentry/geogname im EAD-Namespace vorhanden sind.
* In 'Anreicherung.xlsx' befinden sich jene Informationen, die zur Anreicherung nötig sind: so etwa die Geonames-ID der Gebietskörperschaft für den alle Kind-Orte ausgelesen werden sollen (Zelle B2). Es können auch mehrere Geonames-ID angegeben werden, um etwa bei einem Stadtarchiv neben der Stadt auch den umgebenden Landkreis zu berücksichtigen. Mehrere Geonames-Id müssen nur durch ein „;“ (ohne Leerschlag) getrennt sein.
* Diese Anreicherungsinformationen werden über die geonames-API bezogen. Danach wird über wikidata und den GND-Sparql Endpoint der ZBW Labs versucht, eine geonames - GND Konkordanz zu erzeugen (falls diese nicht im Skript ausgeschalten sind)
* Die so generierten Informationen werden unter data\_geogname zwischengespeichert, damit diese nicht jedes Mal abgefragt werden müssen. Es wird aber trotzdem versucht, GND Identifier, falls etwa der GND Sparql Endpoint Fehlermeldungen zurückgibt, bei einem erneuten Start des Skripts nachzuladen und im Zwischenspeicher abzulegen.
* In 'Anreicherung.xlsx' (Zelle B3) kann eingestellt werden, ob automatisiert alle eindeutig zugeordneten Orte angereichert werden sollen oder nicht. Ansonsten können auch nur jene Orte angereichert werden, die manuell kontrolliert wurden und bei denen in der Spalte Anreicherung im Tabellenblatt 'Treffer' der Wert 'ja' gesetzt wurde. Falls explizit „nein“ gesetzt wird, wird dies in der statistischen Auswertung als falsche Zuordnung gewertet.
* In 'Anreicherung.xlsx' können auch manuell weitere Orte und die entsprechende GND eingetragen werden. Für diese würde dann auch eine Anreicherung stattfinden.
* Falls zwei gleichlautende Ortsnamen existieren, werden diese von der Anreicherung ausgeschlossen und im Tabellenblatt 'nicht eindeutige Treffer' samt der Anzahl ausgegeben.
* Ortsindexbegriffe, die bei der Geonames-Abfrage nicht gefunden werden konnten, werden im Tabellenblatt 'fehlende Treffer' ausgegeben.
* Die den Prozess durchlaufenen EAD-Dateien werden nach derselben Struktur wie im Ordner 'data' unter 'data\_enriched' abgelegt. (falls dies nicht im Skript ausgeschalten ist)
* Es wird, falls vorhanden, die GND unter @authfilenumber eingetragen, ansonsten die Geonames-ID. Unterschieden werden kann der Inhalt über das @source-Attribut.
* Zusätzlich wird eine Csv-Liste unter data\_geogname ausgegeben, die den relativen Pfad der Anreicherung.xlsx-Datei als Namen hat. In dieser sind die Anreicherungsinformationen je Datensatz samt ID aufgelistet, damit diese auf Wunsch auch in die eigene Datenbank zurückgespielt werden können.
* In der Datei Statistik.xls, wir nach EAD-Dateien, VZEs und Indexbegriffen die Anreicherung ausgewertet.

# Einschränkungen

* Wenn das Skript unter Windows läuft, müssen alle Excel-Dateien geschlossen sein, da diese ansonsten gesperrt sind.
* die Groß- und Kleinschreibung ist von Bedeutung ( 'ja', ‚nein‘,'Anreicherung.xlsx')
* die Benennung der Excel-Tabellen darf nicht geändert werden ('Einstellungen', 'Treffer') oder muss im Code nachgeführt werden
* die Felder, in denen die Einstellungen stehen, dürften nicht verändert werden (Zelle B2 für die Start geonames-ID (Gebietskörperschaft), B3 ob Anreicherung manuell oder für alle erfolgen soll ('ja', bzw. anderer Wert (‚nein‘)), B4 ob Anreicherungsinformationen jedes Mal neu von geonames, etc. geholt werden oder aus dem Cache geladen werden sollen ('ja' bzw. anderer Wert))
* alte Werte werden nicht aus dem Anreicherung.xlsx gelöscht, dies muss manuell erfolgen
* momentan erfolgt die Anreicherung um GNDs über den GND Sparql-Endpoint der ZBW Labs, der sehr schnell „außer Atem“ ist, hier müsste wohl ein eigener lokaler SPARQL Endpoint mit den GND Daten aufgebaut werden. Außerdem steht je username nur ein begrenztes freies Kontingent an täglichen Abfragen zur Verfügung **(es sollte, falls das Skript genutzt wird, der username im Skript von marlbaw auch auf einen LABW-Account umgeändert werden)**
* Dadurch, dass vor der Anreicherung alle Kind-Orte eines Geonames Eintrags abgefragt werden, darf diese Gebietskörperschaft nicht zu groß gewählt werden, da ansonsten zu viele Anfragen an Geonames gestellt werden müssten. Die Anreicherung sollte maximal auf kommunaler Ebene (z.B. Kreis Karlsruhe) stattfinden, ansonsten müssten die Geonames-Daten selbst lokal vorgehalten werden. Je größer der Bereich gewählt wird, desto wahrscheinlicher ist es aber auch, dass gleichlautende Ortsnamen existieren.

# Aufbau

**Hauptordner (z. B.: Anreicherung)**

* *enrich\_ead\_geogname\_entries.py:* Skript muss sich immer hier befinden oder die relativen Pfade im Skript müssen angepasst werden (konkret, ohne es ausprobiert zu haben in den folgenden Variablen: ead\_folder, enrichment\_folder, results\_folder)
* **data**
* **flexible Unterordnerstruktur**: es werden nur jene XML-Dateien verarbeitet, bei denen sich im gleichen Ordner eine Anreicherung.xlsx Datei befindet
* *\*/Anreicherung.xlsx*: Informationen, mit welchen Geonames Daten angereichert werden soll, Einstellmöglichkeiten, ob anzureichernde Daten immer neu geholt werden sollen, ob Anreicherung mit oder ohne Kontrolle stattfinden soll
* **data\_enriched**
* **gleiche Ordnerstruktur wie unter data** (wird wo nötig automatisch erzeugt), hier werden die prozessierten EAD-Daten abgelegt (unabhängig davon ob Anreicherungen stattgefunden haben oder nicht)
* **data\_geogname**
* *\*.json* = gesammelte Anreicherungsinhalte von geognames, wikidata, gnd-Sparql-Endpoint, der Name ist die Start-Geonames-ID
* *data\_\*.csv* = Liste aller IDs die angreichert wurden. Aufbau: collection-id, file/item-id, anzureichernder String, Normvokabular (source), normnummer (authfilenumber)
* *Statistik.xlsx:* Auswertung aller Ordner, in denen eine Anreicherung.xlsx gefunden werden konnte.