

Cleaning and enriching a dataset

Schriftprobensammlung des BGBM

Lena-Marie Hoppe

1000 rows
12 columns


- autograph collection comprises digitised manuscript samples spanning more than three centuries
 - however...
 - <https://api.bgbm.org/autographs/v1/list> leads to list of authors, not autographs

overview

- no empty columns
- columns “Geburtsjahr”/”Todesjahr”
 - format should be YYYY, but not standardised
 - other formats: DD.MM.YYYY, D.MM.YYYY, DD.M.YYYY, DD.M. (oder DD.M.)YYYY, D.M.YYYY...
- identifiers: GND number, GUID of Harvard University Index of Botanists (globally unique identifier)
- name separated into last name and first name
- column “Beruf_Tätigkeit”:
 - no standardised abbreviations (und/u.)
 - listing of two or more professions not standardised (either separated with comma or linked with “und”)
- column “Name_andere_Schreibweisen”
 - also used for maiden names of female authors

1887	1961	Dänemark	Botaniker und Algologe
19.6.1881	1971	Costa Rica und Brasilien	Botaniker
10.05.1805	29.03.1877	Deutschland, Berlin	Botaniker, Bryologe
1870	1935	Tansania, Deutschland	Botaniker und Pharmazeut

11.2.1781	11.2.1847	Deutschland	Apotheker und Bryologe
1937	2013	England	
1.10.1806	12.8.1888	Italien	Botaniker
12.01.1831	23.04.1906	Deutschland	Botaniker, Philologe
1924	2007	USA, Deutschland	Botaniker, Bibliothekar
9.5.1809	18.2.1884	USA	Botaniker
8.4.1779	31. [oder 30.] 1.1856	Deutschland	Botaniker
1858	1921	Deutschland	Botaniker, Hochschullehrer

 **Name_andere_Schreibweisen**

Bodelschwingh-Heide

nur Bernard als Vorname

geb. Goujaud

cleaning the data

- header ✓
 - removing „_-result-_-“: manually for every column
 - ”Rename” and “Remove” very close together...
- add column “Geschlecht” ?
- add column “fullname” ✓
 - by joining columns ”Vornamen” and “Name”
- reconciling with WikiData ✓
 - “Name” + (GND + Vornamen) as relevant columns
 - much faster than just Name and better result
 - link to correct entities !!

▼ _ - result - _ - Name

▼ _ - result - _ - Name_andere_Schreibweisen

▼ _ - result - _ - Vornamen

▼ Name

▼ Name_andere_Schreibweisen

▼ Vornamen

▼ fullname

Lujo Adamović

[Choose new match](#)

Adolf Friedrich zu Mecklenburg

[Choose new match](#)

Paul Aellen

[Choose new match](#)

Carlo Allioni

[Choose new match](#)

Arthur Hugh Garfit Alston

[Choose new match](#)

Konstantinos Th. Anagnostidis

[Choose new match](#)

Joseph Anders

[Choose new match](#)

Nils Johan Andersson

[Choose new match](#)