Cleaning and enriching a dataset – UPDATE

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

cleaning & enriching the data

- header ✅
  - removing „_-result-_- “: manually for every column
- added column “fullname” ✅
  - by joining columns ”Vornamen” and “Name”
- added column “Geschlecht” from WikiData ✅
- reconciling with WikiData ✅
  - “Name” + (GND + Vornamen) as relevant columns
  - link to correct entities ✅ (no new items created so far)
- added columns “botanisches Autorenkürzel gemäß IPNI”

cleaning & enriching the data

- added columns “Geburtsdatum“/“Sterbedatum“ ✅
  - imported from WikiData → new columns based off of Geburts- und Sterbedatum:
    - Geburtsdatum_str/Sterbedatum_str: for better readability
    - Geburtsjahr/Sterbejahr: might be superfluous ?
- column “Beruf“ separated into Beruf 01-07 ✅
  - “und“/“u.“ replaced with “,“
- reconciling columns ”Beruf 01-07” !!
- separating and reconciling column “Wirkungsort” !!
- removed “kein Eintrag” from GND/HUH_GUID ✅5

Geburtsdatum	Geburtsdatum_str	Geburtsjahr	Geburtsdatum_org	Sterbedatum	Sterbedatum_str	Sterbejahr	Sterbedatum_org
1864-07-31T00:00:00Z	31.07.1864	1864	1864	1935-07-19T00:00:00Z	19.07.1935	1935	1935
1873-10-10T00:00:00Z	10.10.1873	1873	1873	1969-08-05T00:00:00Z	05.08.1969	1969	1969
1896-05-13T00:00:00Z	13.05.1896	1896	1896	1973-08-19T00:00:00Z	19.08.1973	1973	1973

- original date of birth/data in column "_org"
- problem: only year given → date set to yyyy-01-01