STARTING A HUMANITIES DATABASE WITH HEURIST

Digital Lab
November 15, 2022

Dr. des. Thorben Pelzer
Research Centre Global Dynamics
Leipzig University
Databases | Thorben Pelzer

SOME INSPIRATION

The Virtual Museum of Balinese Painting

Balinese paintings

Homeric citations

Iranian refugees

Australian students

Medieval cuisine

suffrage activists
"CHINESE ENGINEERS RELATIONAL DATABASE" (CERD)

- Republican era (1912–1949)
- 17,472 individual engineers
- 657 colleges
  - 17,159 degrees
- 9,200 workplaces
  - 25,303 employments

- https://home.uni-leipzig.de/cerd/
PRIMARY SOURCE: WHO’S WHO OF CHINESE ENGINEERS (1941)
TODAY’S TALK

- What is a relational database?
- What is Heurist?
- Designing a database
- Managing a database
- Exploring and analysing
DESIGNING A DATABASE

- What are you asking?
  - exploratory (catalogue, all data available)
  - research question (specific use, preselect data)

- What are you collecting?
  - text
  - numbers
  - hierarchical classifications
  - images, video
  - spatial data
  - dates
RELATIONAL DATABASE DESIGN

how not to save your data

<table>
<thead>
<tr>
<th>Name</th>
<th>City</th>
<th>Height</th>
<th>Floors</th>
<th>Year</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silberturm</td>
<td>Frankfurt</td>
<td>166.3</td>
<td>32</td>
<td>1978</td>
<td>Wikipedia</td>
</tr>
<tr>
<td>Post Tower</td>
<td>Bonn, Germany</td>
<td>162.5</td>
<td>42</td>
<td>December 2002</td>
<td>en.wikipedia.org</td>
</tr>
<tr>
<td>Kölnnturm</td>
<td>Cologne</td>
<td>148.1</td>
<td>43</td>
<td>2001/12</td>
<td>Brockhaus S. 23</td>
</tr>
<tr>
<td>Atlantic Hotel</td>
<td>Bremerhaven</td>
<td>147</td>
<td>23</td>
<td>2008</td>
<td>Brockhaus S. 10</td>
</tr>
<tr>
<td>Sail City</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>City-Hochhaus</td>
<td>Leipzig</td>
<td>142</td>
<td>36</td>
<td>1972</td>
<td>Brockhaus</td>
</tr>
<tr>
<td>Colonia-Haus</td>
<td>Cologne, NRW</td>
<td>124.9</td>
<td>32</td>
<td>2001, demolished 2020</td>
<td>?</td>
</tr>
</tbody>
</table>

relational design

<table>
<thead>
<tr>
<th>City ID</th>
<th>Name</th>
<th>State</th>
<th>Country</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Frankfurt</td>
<td>Hessen</td>
<td>Germany</td>
<td>50.1109</td>
<td>8.6821</td>
</tr>
<tr>
<td>2</td>
<td>Bonn</td>
<td>NRW</td>
<td>Germany</td>
<td>50.7374</td>
<td>7.0982</td>
</tr>
<tr>
<td>3</td>
<td>Cologne</td>
<td>NRW</td>
<td>Germany</td>
<td>50.9375</td>
<td>6.9603</td>
</tr>
<tr>
<td>4</td>
<td>Bremerhaven</td>
<td>Bremen</td>
<td>Germany</td>
<td>53.5396</td>
<td>8.5809</td>
</tr>
<tr>
<td>5</td>
<td>Leipzig</td>
<td>Saxony</td>
<td>Germany</td>
<td>51.3397</td>
<td>12.3731</td>
</tr>
</tbody>
</table>

- data not “clean”
- prone to have unintentional doublets
- difficult to maintain/batch edit
WHAT IS HEURIST?

- “solution to the data management needs of humanities”
- “design, populate, explore and publish” databases
- advantages:
  - built specifically with humanities in mind (cf. Access)
  - easy to use (cf. SQL/NoSQL databases)
  - free (cf. FileMaker)
  - online (simultaneous team collaboration)
  - hosting available (or host yourself)
- https://heuristnetwork.org/
TWO SIDES OF HEURIST

- **backend (data input)**
  - accessible to you and your team
  - “design” the database: what fields do you need?
  - populate the database

- **frontend (data exploration)**
  - accessible to anyone
  - build search queries and facets
  - add plugins such as maps, graphs
  - create links for specific entries
BACKEND (INPUT)

1. define data fields early on
2. prepare input forms for you or your team
3. use the forms to populate the database (or batch import data)
4. batch edit data, export selected data, create interfaces
BACKEND

- Additional backend features:
  - helper that checks for potential duplicates
  - website builder for public interface
    - Plugins (map, network, etc.)
  - builder for default “facets”
  - …and much more
FRONTEND (USER INTERFACE)
FRONTEND (USER INTERFACE)
QUERY EXAMPLES
## Databases

<table>
<thead>
<tr>
<th>Thorben Pelzer</th>
</tr>
</thead>
</table>

### SFB 1199

#### Search

<table>
<thead>
<tr>
<th>Name</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>Beijing</td>
</tr>
</tbody>
</table>

#### Results

<table>
<thead>
<tr>
<th>Title</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>China Vehicular Power Plant</td>
<td>Beijing</td>
</tr>
<tr>
<td>National Construction Corporation</td>
<td>Beijing, China</td>
</tr>
</tbody>
</table>

#### Map

- [Location Map](#)
### HYPERLINKED RESULTS

#### Zhang, Yanshao (1900–)

<table>
<thead>
<tr>
<th>Type: Person</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Surname</strong>: Zhang</td>
</tr>
<tr>
<td><strong>Given name</strong>: Yanshao</td>
</tr>
<tr>
<td><strong>Gender</strong>: Male</td>
</tr>
<tr>
<td><strong>Birth year</strong>: 1900</td>
</tr>
<tr>
<td><strong>Native place</strong>: City of Zhangjiang, Jiangsu, China</td>
</tr>
<tr>
<td><strong>Tertiary education</strong>: Study of Zhang Yanshao, 1922: Electrical Engineering, Department of Shanghai Jiaotong University or Peking University, Beijing, China</td>
</tr>
</tbody>
</table>

#### Employment of Zhang Yanshao

- Employment of Zhang Yanshao, 1937–1947: Jardine Matheson, Shanghai, China
- Employment of Zhang Yanshao, 1925–1930: Shanghai Public Utilities Company, Shanghai, China
- Employment of Zhang Yanshao, 1930–1933: Nanking Electric Power Plant, Shanghai, China
- Employment of Zhang Yanshao, 1933–1935: Nanking Electric Power Plant, Shanghai, China
- Employment of Zhang Yanshao, 1937–1947: Nanking Electric Power Plant, Shanghai, China
- Employment of Zhang Yanshao, 1947–1947: Nanking Electric Power Plant, Shanghai, China
- Employment of Zhang Yanshao, 1947–1947: Nanking Electric Power Plant, Shanghai, China
- Employment of Zhang Yanshao, 1947–1947: Nanking Electric Power Plant, Shanghai, China
- Employment of Zhang Yanshao, 1947–1947: Nanking Electric Power Plant, Shanghai, China
- Employment of Zhang Yanshao, 1947–1947: Nanking Electric Power Plant, Shanghai, China
- Employment of Zhang Yanshao, 1947–1947: Nanking Electric Power Plant, Shanghai, China
- Employment of Zhang Yanshao, 1947–1947: Nanking Electric Power Plant, Shanghai, China
QUERY BUILDER: MORE COMPLEX CONDITIONS

“Show me all individuals with surname WANG born BEFORE 1920 who graduated in something OTHER THAN civil engineering in EITHER Germany OR France and later worked at a RAILWAY

```json
["t":"58"],{"f:997":"-9483"}, {"linkedfrom:940":[{"t":"10"}, {"f:1":"wang"}],"f:10":"<=1920"}, {"linked_to:942":[{"t":"61"},{"linked_to:21":[{"t":"4"}, {"f:1":"鐵路"}]})}], {"linked_to:964": [{"t":"57"},{"linked_to:238": [{"t":"12"},{"f:26":"182,110"}]})}],{"sortby":"t"}]
```
USING DATABASES IN RESEARCH

- input and manage
- explore and analyse
- process
- visualisations

Databases

- HEURIST
- Zenodo
- R Studio
- Tableau
- Gephi.org
- QGIS
- ArcGIS
- R AWGraphs

data

database

repository

statistics

networks

spatiality
NETWORK ANALYSIS
SPATIAL ANALYSIS

Pull of selected colleges, graduates up to 1937

Push of selected colleges, occupations up to 1937

Workplace mobility of engineers born in selected provinces, by percentile
VISUALISING TEMPORALITY

New Appointments of Chinese Engineers, 1910–1940

Data set has late 1930s bias. 53% of appointments are excluded as no date given.

Data Source CC-BY 4.0: Thorben Pelzer et al. (2021): Chinese Engineers Relational Database.
Basemaps: CHGIS © Fairbank Center & Fudan University, Historical Basemaps GPL 3.0, André Ourednik, US History Maps CC-BY-SA 3.0, Michael Porath.
CHALLENGES

- when using historic sources:
  - bad print, bad OCR, factual errors (edit or keep?)
- fragmentariness, granularity, ambivalence
  - historic name changes, change of administration
  - geodata
- avoid creating doublets
  - avoid false mergers
DATA LINKING

- link entries across datasets
- case: link biographies in Chinese Engineers Relational Database (CERD) and Modern China Biographical Database (MCBD)
- issue to resolve: how to ensure two individuals are the same person?
  - what criteria need to be fulfilled to argue two persons are “the same”?
  - can we develop an automated check?
  - can we move from linking to merging without creating doublets?
QUESTIONS