UNIVERSITÄT
LEIPZIG

Leipzig Research Centre Global Dynamics

Thorben Pelzer

Chinese Engineers Relational Database (CERD) Design Manual

RECENTGLOBE

WORKING PAPER SERIES
No. 44

UNIVERSITÄT
LEIPZIG

Leipzig Research Centre Global Dynamics

IMPRINT

This working paper is part of the Working Paper Series of the Collaborative Research Centre (SFB) 1199 “Processes of Spatialization under the Global Condition”.

This working paper is also part of the Working Paper Series of ReCentGlobe, to which the SFB 1199 contributes since 2020.

Pelzer, Thorben. “Chinese Engineers Relational Database (CERD) Design Manual”. ReCentGlobe Working Paper 44. Leipzig University, 2020.

This ReCentGlobe Working Paper is published online and can be downloaded from the website:

https://www.recentglobe.uni-leipzig.de/zentrum/detailansicht/artikel/chinese-engineers-relational-database-cerd-design-manual-2020-09-28/

Leipzig Research Centre Global Dynamics (ReCentGlobe)
Universität Leipzig

Visitor’s Address:

Nikolaistrasse 6–10
Strohsackpassage, 5th floor
D-04109 Leipzig

Tel.: 0049 / (0)341 / 97 37 884
E-Mail: recentglobe@uni-leipzig.de
www.recentglobe.uni-leipzig.de

Postal Address:

ReCentGlobe
Universität Leipzig
IPF 348001
D-04081 Leipzig

SFB 1199

Processes of Spatialization under the Global Condition

Thorben Pelzer

Chinese Engineers Relational Database (CERD) Design Manual

Working paper series des SFB 1199 an der Universität Leipzig
Nr. 21

Collaborative Research Centre (SFB) 1199
„Processes of Spatialization under the Global Condition“
at Leipzig University

Funded by

Deutsche Forschungsgemeinschaft

Thorben Pelzer
Chinese Engineers Relational Database (CERD) Design Manual

This working paper is part of the Working Paper Series of the Collaborative Research Centre (SFB) 1199 "Processes of Spatialization under the Global Condition". This working paper is also part of the Working Paper Series of ReCentGlobe, to which the SFB 1199 contributes since 2020.

09 / 2020

Vertrieb:
Leipziger Universitätsverlag GmbH, Oststraße 41, 04317 Leipzig
info@univerlag-leipzig.de

ISBN: 978-3-96023-220-9
ISSN: 2510-4845

Chinese Engineers Relational Database (CERD) Design Manual

Thorben Pelzer

Content

Abstract

The CERD database is a register of engineers from the Chinese Republican period (1912–1949). The digital analysis of historical registers has led to a prosopographic catalogue of persons, educational institutions, and companies. The data can be put in relation to one another by researchers to answer individual research questions. This paper describes the architecture of the database, explains how it works, and presents ways of analysing the data.

Introduction

The following manual describes the aims, structure, and operation of the Chinese Engineering Relational Database (CERD). The database is the quantitative part of the broader collaborative research group "Chinese Engineers and their Spatial Imaginations: Architects of an Interconnected Nation, 1906–1937", led by Elisabeth Kaske.[1] The subproject is situated within the Collaborative Research Centre (SFB) 1199, which investigates "Processes of Spatialization under the Global Condition".[2]

The subproject aims to analyse the relationship between infrastructure and the nation-state. It investigates the introduction of the nation-state as a spatial format, the infrastructural formatting of the "Chinese" territory, and the infrastructural dimension of integrating China within a global spatial order. The group will look at three different groups of actors: the Ministry of Communications *(jiaotong bu),* polytechnic schools and universities, and professional engineering associations. The database serves as an aid for the subproject as a whole and connects the different angles, as it comprises data relevant to all the research questions.

Aim and Scope

The database's origin and function within the research centre are reflected in its structure: the database revolves around individuals (actors being the research object of the research centre subgroup) as well as locations (spatialities being the research object of the overall research centre).

CERD is not a general-purpose database like the China Biographical Database (CBDB), a vast project that collects a wide range of biographical detail, with a focus on the pre-modern period.[3] In contrast, data collection for CERD is designed with the specific goals of the research project in mind. Therefore, the database features a deliberately limited amount of defined entities.

The database serves as a tool of quantitative exploration and as a support for qualitative studies. This primarily relates to two questions within the spatialization research complex. The first is that of mobilities and the second is that of networks. Through the uncovering of mobilities, such as the extent of foreign tertiary education or the flows of internal workplace migration, data analysis indicates potentials of spatial literacy as well as provides a ground to observe spatial formatting as a performance of movement in space. Meanwhile, individual and organizational networks unearthed through the dataset give insight into the human capital and social connectedness *(guanxi)* of the engineers, which were prerequisites for their agency as infrastructural entrepreneurs and which may also have worked as influences on the direction of the work they produced. In all aforementioned cases, theories of space are less presupposed into the database design than they are part of the data output and its interpretation.

To this regard, while contrived within a research context of the sociology of space, third parties are invited to analyse the data for their aims as well. To enable data analysis outside the entity spectrum of CERD, the complete or selected data can be exported and integrated into other databases, as described in chapter 4. The data exportability is ensured with related investigations such as the "Elites, Networks and Power" research project in mind.[4]

1 The author of this manual is also a researcher within the group. Chen Hailian is co-opted to the SFB research group, and funded through her BMBF project "Die Wegbereiter von Chinas Aufstieg zur Technologiemacht: Technische Bildungseinrichtungen und ihre Studierende im Zeitalter des Globalen Wandels, 1860–1911" [The pioneers of China's rise to technological power: Technical educational institutions and their students in the age of global transformation, 1860–1911]. The following diligent student assistants pioneered our data input: Oliver Dieckmann, William Chin Fung Ng, and Felix Opper.

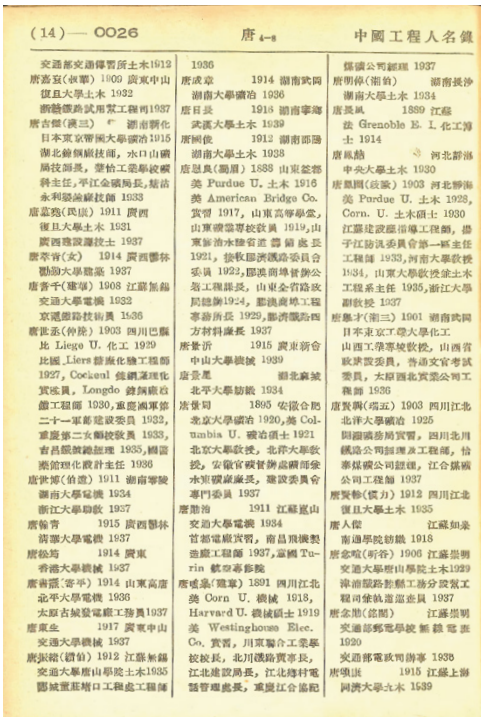
2 Collaborative Research Centre 1199, "Research Programme", https://research.uni-leipzig.de/~sfb1199/research/research-programme/. (accessed 4 May 2020).

3 Harvard University, Academia Sinica, and Peking University, "China Biographical Database", 24 April 2018, https://projects.iq.harvard.edu/cbdb.

4 C. Henriot, "Elites, Networks, and Power in Modern Urban China (1830–1949): Historical 'Big Data' in Modern Chinese History", June 2018, https://enepchina.hypotheses.org/files/2018/06/ERC_Project-description.pdf.

Sources

The data mining process can be divided into three phases according to the priority in the project and the effort required to collect the data from the sources. Where automatized digitalization is possible in an efficient manner, the sources have been scanned using optical character recognition (Tesseract OCR) and the results checked twice before finally being entered into the system. Minor sources of smaller scope are instead collected manually (see below).


(14)— 0026 唐 4–8 中國工程人名錄

交通部交通傳習所土木1912
唐嘉宜(叔華) 1909 廣東中山
復旦大學土木 1932
浙贛鐵路試用幫工程司1937
唐吉樑(漢三) 湖南醴陵
日本東京帝國大學礦冶1915
湖北鋼鐵廠技師，水口山礦
局技師長，華怡工業學校礦
務主任，平江金礦局長，楚古
永利鹼廠廠技師 1933
唐嘉逸(民康) 1911 廣西
復旦大學土木 1931
廣西電燈廠技士 1937
唐萃青(女) 1914 廣西鬱林
鬱江大學肄業 1937
唐書年(建章) 1908 江蘇無錫
交通大學電機 1932
京滬鐵路技術員 1936
唐世孟(仲陵) 1903 四川巴縣
比 Liege U. 化工 1929
比國 Liers 發電化驗工程師
1927，Cockeul 鐵網溶理化
實驗員，Longdo 鋼鐵廠造
鑑工程師 1930，重慶國軍第
二十一軍部建設委員 1932，
重慶第二女師校教員 1933，
吉昌鐵廠總經理 1935，國防
廠館理化設計主任 1936
唐世郊(伯澄) 1911 湖南寧鄉
湖南大學電機 1934
浙江大學助教 1937
唐愉青 1915 廣西鬱林
廣州大學電機 1937
唐松筠 1914 廣東
香港大學機械 1937
唐南孺(寄平) 1914 山東高唐
北平大學電機 1936
太原古城發電廠工務員1937
唐東生 1917 廣東中山
交通大學機械 1937
唐振緒(績伯) 1912 江蘇無錫
交通大學唐山學院土木1935
鄂城電廠漢口工程處工程師
1936
唐成章 1914 湖南武岡
湖南大學礦冶 1936
唐日長 1916 湖南寧鄉
武漢大學土木 1939
唐國俊 1912 湖南邵陽
湖南大學土木 1938
唐慈良(湘眉) 1888 山東益都
美 Purdue U. 土木 1916
美 American Bridge Co.
實習 1917，山東高等學堂，
山東礦業專校教員 1919，山
東旅治水勘官差 營員處長
1921，徐州津浦路委員會
委員 1923，鄂漢商埠督辦公
署工程課長，山東全省路政
局總辦1924，鄂漢商埠工程
事務所長 1929，鄂漢鐵路西
方材料廠長 1937
唐曾沂 1915 廣東新會
中山大學機械 1939
唐登雲 湖北廣濟
北平大學肄業 1934
唐登司 1895 安徽合肥
北京大學礦冶 1920，美 Col-
umbia U. 碩冶碩士 1921
北京大學教授，北洋大學教
授，安徽官礦督辦處礦師兼
水口礦廠廠長，建設委員會
專門委員 1937
唐陶治 1911 江蘇崑山
交通大學電機 1934
首都電廠實習，南昌飛機製
造廠工程師 1937，廬國 Tu-
rin 航空專修院
唐咸豪(雅軍) 1891 四川江北
美 Corn U. 機械 1918，
Harvard U. 機械碩士 1919
美 Westinghouse Elec.
Co. 實習，川東聯合工業學
校校長，北川鐵路實事長，
江北建設局長，江北縣村電
話管理處長，重慶江合鹽記
煤礦公司總經理 1937
唐明仲(湘伯) 湖南長沙
湖南大學土木 1934
唐長風 1889 江蘇
法 Grenoble E. I. 化工博
士 1914
唐風鳴 河北靜海
中央大學土木 1930
唐鳳圖(啟愚) 1903 河北靜海
美 Purdue U. 土木 1928，
Corn. U. 土木碩士 1930
江蘇建設廳淮揚工程師，揚
子江防汛委員會第一區主任
工程師 1933，河南大學教授
1934，山東大學教授兼土木
工程系主任 1935，浙江大學
副教授 1937
唐鳳才(湘三) 1901 湖南武岡
日本東京工學大學化工
山西工學專校教授，山西省
政府設委員，省通文官考試
委員，太原西北實業公司工
務師 1936
唐賢興(瑞五) 1903 四川江北
北洋大學礦冶 1925
四川鐵路局實習，四川北川
鐵路公司經理及工程師，怡
泰煤礦公司經理，江合煤礦
公司工程師 1937
唐賢修(儀力) 1912 四川江北
復旦大學土木 1935
唐人傑 江蘇如皋
南通學院紡織 1918
唐念哉(昕谷) 1906 江蘇崇明
交通大學唐山學院土木1929
滬南鐵路勘測工務分設幫工
程司兼軌道處主任 1937
唐念慈(慈閣) 江蘇崇明
交通部郵電學校無線電班
1920
交通部電政司技事 1936
唐頌康 1915 江蘇上海
同濟大學土木 1939

Fig. 1. A typical page of the Gongchenren minglu

Phase 1 – Gongchengren minglu[5]

The first set of data being inputted is the *Gongchengren minglu*, a 1941 directory of engineers published by the National Resources Commission *(Guojia ziyuan weiyuanhui*, NRC*)*.[6] With approximately 13,000 entries, this directory is the largest of its kind. As the directory only concerns engineers, and as it was published four years after the end of the researched period, a manual sorting of the data is not necessary. The bullet point style of entries further eases the flow of data mining (see figure 1).

Phase 2 – Further directories

In the second phase, the data will be supplemented using further directories available. In contrast to the *Gongchengren minglu*, these directories are not exclusively listing engineers, but they include many of them. These sources include *Who's who* indices that list influential individuals[7] and historical directories of students.[8] Secondary literature further complements the results, listing students,[9] technocrats,[10] or influential individuals in general.[11]

Phase 3 – Various archival sources

The data mining of the last phase is expected to be the least efficient. The sources consist of local gazetteers *(difang zhi)*, member lists of professional engineering associations, and other scattered archival materials. As these sources have no consistent structure, they need to be worked through on

5 Zīyuán wěiyuánhuì 資源委員會 [National Resources Commission], Zhōngguó gōngchéngrén mínglù: 中國工程人名錄 [Who's Who of Chinese Engineers], Beijing: Shāngwù Yìnshūguǎn 商務印書館, 1941.

6 We thank Elisabeth Köll for providing us with a copy of the *Gongchengren minglu*.

7 A. Ramsay (ed.), *The Peking Who's Who*, Beijing: The Tientsin Press, 1922; M. C. Powell (ed.), *Who's Who in China*, Shanghai: The China Weekly Review, 1925; The China Weekly Review, *Who's Who in China· Biographies of Chinese Leaders*, 5th ed., Shanghai: The China Weekly Review, 1936.

8 *The Chinese Students' Directory* (s. l.: Chinese Students' Alliance, 1914); Běijīng Qīnghuá xuéxiào 北京清華學校, *Yóu Měi tóngxué lù·* 游美同學錄 [Who's who of American returned students] (Beijing: Qīnghuá xuéxiào 清華學校, 1917); H. Tung (ed.), *Who's Who of the Chinese Students in America*, Berkeley: Lederer, Street & Zeus, 1921; China Institute in America, *Theses and Dissertations· By Chinese Students in America*, New York: China Institute in America, 1927; E. C. Koo [Gù, Yùzhēn 顧毓珍], *Chinese Students Directory· For the Past Fifty Years*, Cambridge: Massachusetts Institute of Technology Chinese Students Directory, 1931.

9 T. Yuan [Yuán, Tónglǐ 袁同禮], *A Guide to Doctoral Dissertations by Chinese Students in America· 1905–1960*, Washington, DC: Sino-American Cultural Society, 1961; T. Yuan [Yuán, Tónglǐ 袁同禮], *Doctoral Dissertations by Chinese Students in Great Britain and Northern Ireland· 1916–1961* (s. l.: s. n., 1963); T. Yuan [Yuán, Tónglǐ 袁同禮], *A Guide to Doctoral Dissertations by Chinese students in Continental Europe· 1907–1962*, Washington, D.C.: Sino-American Cultural Society, 1964.

10 P. Zhāng 张朋园 and H. Shěn 沈怀玉 (eds.), *Guómínzhèngfǔ zhíguān niánbiǎo 1925–1949·* 国民政府职官年表 *(1925–1949)* [Annals of the nationalist government office holders, 1925–1949], Taipei: Academia Sinica, 1987; S. Liú 劉壽林 et al. (eds.), *Mínguó zhíguān niánbiǎo·* 民國職官年表 [Republican yearbook of office holders], Beijing: Zhōnghuá shūjú chūbǎn 中華書局出版, 1995.

11 H. L. Boorman and R. C. Howard (eds.), *Biographical Dictionary of Republican China*, New York: Columbia University Press, 1967.

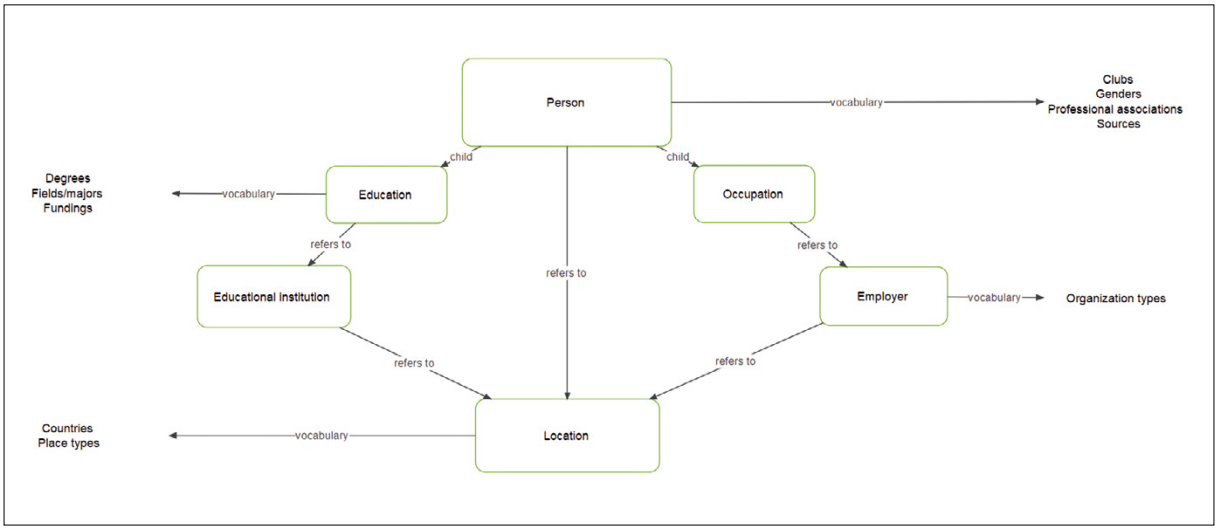


Fig. 2. CERD structure schematic

a page-by-page basis, and as they are often of badly preserved condition and have a poor print quality, the information is faster inserted by hand than through OCR.

Structure

The database centres on biographical entries, the so-called person records (see figure 2). Each person record can have an unlimited amount of child records. These detail the person's career path, either in the form of an educational record or an occupational record. These child records use their own predefined tables, but they cannot exist without referring to a person as their parent record. They describe an individual's temporarily limited education or employment, not the place or location where this education or employment took place.

The educational and occupational child records refer to places where the person studied or worked. These are typically schools or companies. What differentiates these places from the child records is that each such institution exists as a stand-alone record within the system, without a unique parent record. In this way, many different persons' careers can refer to the same institutional record. Instead of using a person record as the start of an enquiry, researchers can also start by browsing a school's or a company's record and from there track which individuals were associated with that institution.

Apart from the child records, the three table types – person, educational institution, and employer – all refer to a location record. For the individual, this is their birthplace, while in the other cases, this is place where the described entry – the headquarters or a subbranch – was located. There is only one location record type, regardless of whether the reference originates from a person, a school, or an employer. The location records are also self-referential, as every Chinese location that is not a province is itself referring to a provincial record (which is also a record that uses the location table).

Apart from (1) entries that are unique to one record (such as a person's name or the name of a company) and (2) entries that refer to other tables (such as a person's birthplace entry referring to a specific location table), the third entry type is the reference to a dictionary. Within the context of CERD, dictionaries are lists of precompiled vocabularies that have no additional information assigned to them. They are used to avoid ambivalence, differences in spelling, etc. of word entries (varchars). These are, for example, lists of company types to tag employers and lists of administrative division types to tag locations. These vocabulary lists can organically grow as the data input advances. However, the vo-

cabularies themselves are incapable of receiving additional information (i.e. additional columns), which differentiates the dictionaries from full-fledged record tables.

Both the six record table types (see figure 2) and the vocabulary list entries have unique IDs automatically assigned to them. These IDs can be used for queries (see chapter 3). In the case of the tables, the IDs are listed in the operator columns below. In the case of the vocabulary lists, as these dynamically grow, the design manual refrains from listing them with their corresponding IDs, although these can be discovered by the tech-savvy using the "defterms" table of the offline SQL export (see chapter 4.2).

Primary Record: The Person Table

Person tables are the primary record type of the database. They only include the most basic information about an individual – their name(s), lifespan, place of birth, gender, as well as a list of memberships in various organizations (see below). The details of their career are instead given in child records that are linked to the primary record (see chapter 2.2).

While the online interface prioritizes the *Hanyu pinyin* transcription of a name, this romanization system was not developed until the 1950s. Instead, when engineers adopted an international spelling of their name, they chose their own means of transcription. They sometimes transcribed their names according to the postal romanization system, but they often took many liberties. In CERD, this kind of spelling is referred to as a person's "conventional" name. In the case that sources list further names for a person, for example a courtesy name (*zi*) or additional conventional names, these are included as "alternate names".

Description	Type	Value	Priority	Operator
ID	integer	unique	automatic	
Sources	dictionary → sources	multi	required	f:996
Family name, *Hanyu pinyin*	varchar	single	required	f:1
Given name, *Hanyu pinyin*	varchar	single	recommended	f:18
Chinese characters	varchar	single	recommended	f:993
Family name, conventional	varchar	single	optional	f:994
Given name, conventional	varchar	single	optional	f:995
Alternate names	varchar	multi	optional	f:132
Gender	dictionary → genders	single	optional	f:20
Birth date	integer	single	recommended	f:936
Place of birth	record → location	single	recommended	f:134
Death date	integer	single	recommended	f:937
Secondary & preparatory education	child → education	multi	optional	f:939
Tertiary education	child → education	multi	recommended	f:940
Occupations	child → occupation	multi	recommended	f:942

Description	Type	Value	Priority	Operator
Professional associations	dictionary → associations	multi	optional	f:1000
Student & alumni clubs, fraternities, non-political service organizations	dictionary → clubs	multi	optional	f:999
Related persons	record → person	multi	optional	
Photograph	digital image file	multi	optional	

Clubs

The database knows two forms of membership: the membership in professional engineering organizations and the membership in clubs. The latter pertains to all forms of membership not included in the former and that are not a form of paid regular labour. This includes membership in student and alumni organizations, fraternities, and non-political service organizations.

Genders

The system knows the genders female, male, other, and unknown. In the case where historical dictionaries distinctively mark women, all other individuals are assumed male.

Sources

The system does not allow for the creation of a new primary record without the reference to a proper source. Source fields are multi-value entries so that existing records can at any time be supplemented with additional information gathered from a variety of sources.

Professional associations

The organization of engineers in professional associations is one of the primary research interests of the research project. The list includes any organization that defines its common membership through the occupation of engineering or its subdisciplines. Some of the associations can be suborganizations or regional branches of others, but an engineer is only linked to the branch or level of an organization as is recorded in the documented biographical information and historical membership lists.

Relationship types

Professional relations are an outcome of data analysis and must not be added manually. The relationship entry is reserved for family relations (child, parent, or sibling) and mentorships (student or teacher). As relationships must be linked to existing person records, only relations to engineers are part of the database.

Child records

CERD features two child record tables, which are similar in form and function. They detail the education(s) and occupation(s) of a person. Theoretically, child records have their own ID and can be accessed and displayed on their own. However, they cannot be created without an existing parent record attached to them, which differentiates them from the secondary records (see chapter 2.3).

Education

Each person can have an unlimited amount of education records assigned to them. While the education record can be assigned to a person record as either a (a) secondary and preparatory, or (b) a tertiary form of education, both entries use the same child record mask.

Description	Type	Value	Priority	Operator
ID	integer	unique	automatic	
Parent record	record → person	single	required	f:247
Educational institution	record → educational institution	single	recommended	f:964
Fields/majors	dictionary → fields/majors	multi	required	f:997
Degree	dictionary → degrees	multi	recommended	f:969
Start date	integer	single	optional	f:10
Graduation date	integer	single	recommended	f:9
Funding	dictionary → funding	single	optional	f:998

Degrees

For the case that the source specifies the degree awarded, this list includes titles from different academic cultures, such as bachelor, master, doctor, diploma, and *juren*.

Fields/majors

The fields of study include general studies such as "engineering" and "sciences" but also more specialized majors such as chemical, hydraulic, and mechanical engineering. The database also includes entry options for fields only partly related to engineering – for example, administration – and options for degrees completely unrelated to engineering – for example, political sciences – so that a complete educational biography can be inputted, including double degrees and studies pursued before the commencement of an engineering career.

Funding

This entity is aimed at distinguishing between private and government students, especially for the case of studies overseas and the enrolment in preparatory schools. The most prominent subcategory of government-funded students were those funded through the US Indemnity Scholarship Fund.

Occupation

Each person can have an unlimited amount of occupations assigned to them. Occupations are compensated forms of regular labour, which differentiates them from memberships in associations and service organizations.

Description	Type	Value	Priority	Operator
ID	integer	unique	automatic	
Parent record	record → person	single	required	f:247
Employer	record → employer	single	required	f:21
Original occupational description	varchar	multi	recommended	f:983
Start date	integer	single	recommended	f:10
End date	integer	single	recommended	f:11

Secondary Records

The following three record types exist independent of the person record table. The first two – the employer and the educational institution records – are referred to by a person's child records. The third one, the location record, is referred to by both the person record and by all secondary records.

Employer

The employer record provides basic information about a place of work, which can be referred to by the occupation child record. The employer record must not be confused with the dictionary for professional associations or the one for fraternities and non-political service organizations, which are used in the case of uncompensated memberships (see chapter 2.1).

Description	Type	Value	Priority	Operator
ID	integer	unique	automatic	
Name of employer	varchar	single	required	f:1
Short name/acronym	varchar	single	optional	f:2
Organization type	dictionary → organization types	multi	recommended	f:22
Location	record → location	single	recommended	f:134
Mother organization	record → employer	single	optional	f:1003

Organization types

This list defines organization records as public or privately operated. It can further specify them as belonging to a branch such as railroads or government offices.
Mother organization

In cases where a subbranch of an organization is of considerable size, occupied with extraordinary tasks, or located at a different place, these branches are created as their own entries and are hierarchically linked to their mother organization.

Educational institution

The educational institution record template is the shortest within CERD. It only gives the name of the institution and links it with a location record.

Description	Type	Value	Priority	Operator
ID	integer	unique	automatic	
Name of institution	varchar	single	required	f:1
Place	record → location	single	recommended	f:238

Location

The person, employer, and educational institution records, as well as the location record itself, can refer to this entity. For practical reasons, CERD only uses points (longitude and latitude) to refer to geographical locations. For larger geographical entities such as provinces and states, the political centre – e.g. the provincial or state capital – is used. Depending on the situation, users may want to replace these geospatial coordinates with vector shapefiles when creating their own graphical presentations.

Description	Type	Value	Priority	Operator
ID	integer	unique	automatic	
Place type	dictionary → place types	single	recommended	f:113
Place name	varchar	single	required	f:1
Province	record → location	single	recommended	f:1002
Country	dictionary → countries	multi	recommended	f:26
Location	geospatial coordinates	single	recommended	f:28_lat and f:28_long

Countries

A list of (contemporary) countries to give a rough idea of national belonging of the locations; locations can be part of multiple countries for cases such as Republican-era Mongolia. The country entity does not aim to reconstruct the frequently changing political belonging of territories to the various Chinese and foreign political entities.

Place types

Location records have a place type assigned to them. The vocabulary list includes the administrative divisions of the Beiyang Republic (counties, towns, circuits, provinces, areas, regions, prefectures, special administrative regions, and commercial regions) and additionally has an entry for foreign independent countries and for US states. The status of a place may have changed through time, as did the structure of administrative divisions, so this entry can only give a basic idea of hierarchy in relation to other locations.

Online Access

The dataset can be accessed using Heurist,[12] a web database interface that provides both a backend and a frontend for relational research data in the humanities. The CERD web frontend is kindly hosted by the University of Sydney and can be accessed through the following link:

https://heuristplus.sydney.edu.au/heurist/?db=tp86k_sfbdem&website#

The same website is accessible using this shortened URL:

https://tinyurl.com/cerd-leipzig

The online database can be cited, for example, as the following:[13]

T. Pelzer et al. (eds), Chinese Engineers Relational Database. https://heuristplus.sydney.edu.au/heurist/?db=tp86k_sfbdem&website# (accessed 4 May 2020).

URLs for individual records can be created using the "cite as HTML" command within Heurist and can be cited as in the following example:

"Fang, Chuanliu 方傳流 (1914–)", Chinese Engineers Relational Database, T. Pelzer et al. (eds.), https://heuristplus.sydney.edu.au/heurist/?recID=1844&fmt=html&db=tp86k_sfbdem (accessed 4 May 2020).

The Heurist interface can be used to query the database (see figure 3). There are mainly two reasons to do this. The first is historical lookups, for example to find all engineers that worked at a specific company or to find out the educational biography of a person of interest. The other usage case would be to select a limited deataset for further external analysis. For example, a query could select all people born in Hunan in order to export this specific set of data for visualization.

Below, the operators and symbols used for these queries are described, before exemplary queries are presented in order to give a clearer impression of the required syntax.

12 I. Johnson and A. Osmakov, "Heurist Academic Knowledge Management System", Faculty of Arts and Social Sciences, University of Sydney, http://heuristnetwork.org/. (accessed 4 May 2020).

13 A thorough list of contributors is available on the database website. As the list of individuals involved, especially of the assistants inputting data, is constantly growing, it is not feasible to include a complete list of editors in print.

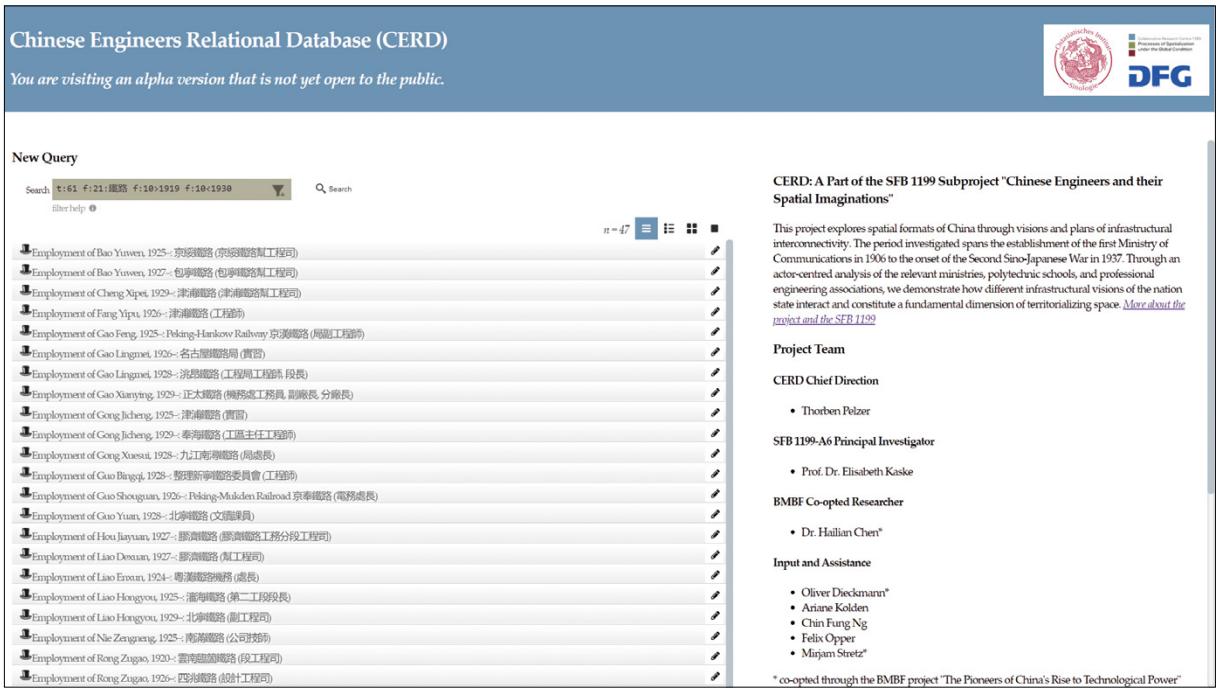


Fig. 3. CERD frontend with sample query

Operators and Symbols

The t-operator ("table") limits queries to specific record tables using their corresponding record ID:

Record table	Operator
Employer	t:4
Location	t:12
Person	t:10
Educational institution	t:57
Occupation	t:61
Tertiary education	t:58
Secondary or preparatory education	t:56

Each field of a table can be addressed using the f-operator ("field") and the corresponding field ID as listed in chapter 2 of this manual.

Tables and fields can be limited using the following symbols:

Symbol	Result
:	includes
=	is exactly
>	greater than (integer)
<	lesser than (integer)
OR	include either
-	omit

The online database distinguishes between simplified and traditional Chinese characters and only uses the latter. Using simplified characters in the search will not yield correct results.

Sample Queries

A fuzzy search that outputs all entities (person, educations, and occupations) with the names "Wang" and "Jingchun" involved:

Wang Jingchun

A more eloquent search that only outputs the primary record (the biographical entity) of individuals named Wang Jingchun:

t:10 Wang Jingchun

These queries are more precise but yield the same results:

t:10 f:1:Wang f:18:Jingchun

t:10 f:993:王景春

t:10 f:994=Wang f:995=Ching-chun

This query shows all engineers with degrees in civil engineering *or* hydraulic engineering:

t:10 f:940:civil OR f:940:hydraulic

To compare, a query that shows only engineers with degrees in *both* civil engineering *and* hydraulic engineering:

t:10 f:940:civil f:940:hydraulic

All engineers who studied at a university in Berlin:

t:10 f:940:berlin

Civil engineering degrees awarded at universities in Shanghai:

土木 f:964:上海

A search for degrees awarded at Tsinghua University that are *not* electrical engineering:

f:964:清華 -electrical

This query outputs all degrees awarded *before* 1912:

t:58 f:9<1912

This query outputs all educational institutions in Guangdong:

t:57 f:238:Guangdong

This query displays every employment at companies with *tielu* ("railway") in their name, limited to hiring in the 1920s:

t:61 f:21:鐵路 f:10>1919 f:10<1930

Exports

While the online frontend can be used for lookups and basic research, more complex analysis will require the export of selected data. The different means of export are described below. It is strongly suggested to use the export function of Heurist for this, but access to offline backups is also granted to secure long-term preservation of the research data. Following the description of different export methods, some usage examples are presented.

Heurist Exports

The backend of Heurist supports the export of queries as comma-separated value (CSV) files. As this is currently not possible through the frontend user interface, the interested researcher will need to either (a) need to inquire for a guest access to the CERD backend or (b) needs to download an offline export and import it into their own Heurist database (see chapter 4.2).

The advantage of using Heurist file exports instead of interpreting the offline backups through other means (see chapter 4.2) is that the files will be easy to read, tables that use data from multiple related record types can be generated easily, and much less computer expertise is necessary in the process.

Offline Backups

Bimonthly exports of the data will be made available through a public data repository once the first phase of data input (see chapter 1.2) has been completed. In contrast to the online version, the offline export does not come with an end-user interface. It is distributed as an XML and as a SQL file to be interpreted locally using the desired analysis tools. The export can be cited as:

Pelzer, T. et al., eds. Chinese Engineers Relational Database (offline export of January 2021).

For most purposes, the XML file will suffice. The XML file follows the Heurist Markup Language (HML) schema, so that it can be imported into other Heurist installations. For other uses, the XML file should first be converted to a CSV, JSON, or Excel file using amply available online convertors. As the table includes the data of all record table types, it must be cleaned manually to only include the types of data required. For instance, deleting all rows but those with the type ID 10 will yield a complete table of all person records (see chapter 3.1).
The second method – using the SQL backup – requires some basic knowledge of SQL. The database file (UTF-8 encoding) can be run on a (virtual) MySQL server in order to export selected data as tables, e.g. in Excel, JSON, or CSV format. The relevant records are in the "records" and "recdetails" tables. The SQL queries will require the same IDs as used for the online queries and as listed in chapter 2.

Usage Examples

After the desired data has been queried and exported, it can be analysed and visualized using tools and libraries such as amCharts, Gephi, RAWGraphs, or Vistorian.[14] The following usage examples are not meant to represent outcomes of the research project. The visualizations are generated using a very early, limited data set that is not a representative account of the prosopography the research eventually aims for. They are included here only as proofs of concept and should not be cited for any research purposes.

The first example (see figure 4) uses data queried for engineers that studied abroad. The visualization traces their career path from their birth province to their educational institution and further to their place of employment. Such visualization can be used to find and display regional preferences to enrol in one or another university or to reveal tendencies for graduates of one university to join the same company.

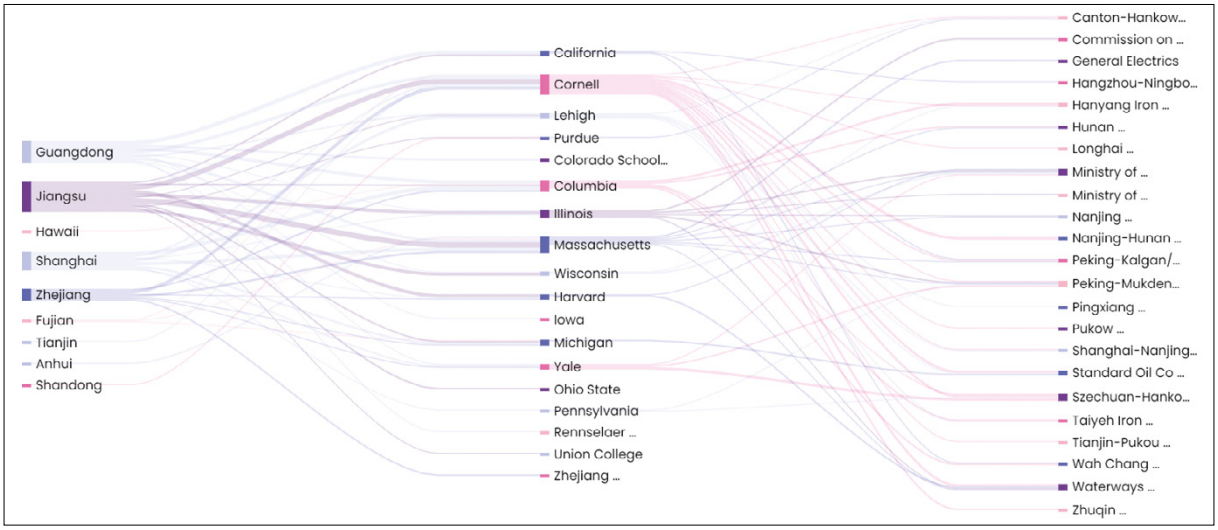


Fig. 4. Visualization of career paths from native place to university to workplace using amCharts

Figure 5 uses the same data selection. The visualization was generated with Palladio.[15] It uses the geospatial data of the engineers' birthplaces and the locations of their educational institutions and connects them accordingly. The marker size is proportional to the occurrence of each birthplace or school.

14 AmCharts, "amCharts: JavaScript Charts & Maps", https://www.amcharts.com/ (accessed 6 May 2020); B. Mathieu, S. Heymann, and J. Mathieu, "Gephi: An Open Source Software for Exploring and Manipulating Networks" (International AAAI Conference on Weblogs and Social Media, 2009); M. Mauri et al., "RAWGraphs: A Visualisation Platform to Create Open Outputs", CHItaly '17: Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter, 2017, https://doi.org/10.1145/3125571.3125585; Bach, B. et al., "NetworkCube: Bringing Dynamic Network Visualizations to Domain Scientists" (Posters of the Conference on Information, 2015), https://hal.inria.fr/hal-01205822. (accessed 6 May 2020).

15 D. Edelstein, N. Coleman, and E. Jewett, "Palladio: Stanford University Humanities + Design", https://hdlab.stanford.edu/palladio/. (accessed 6 May 2020).

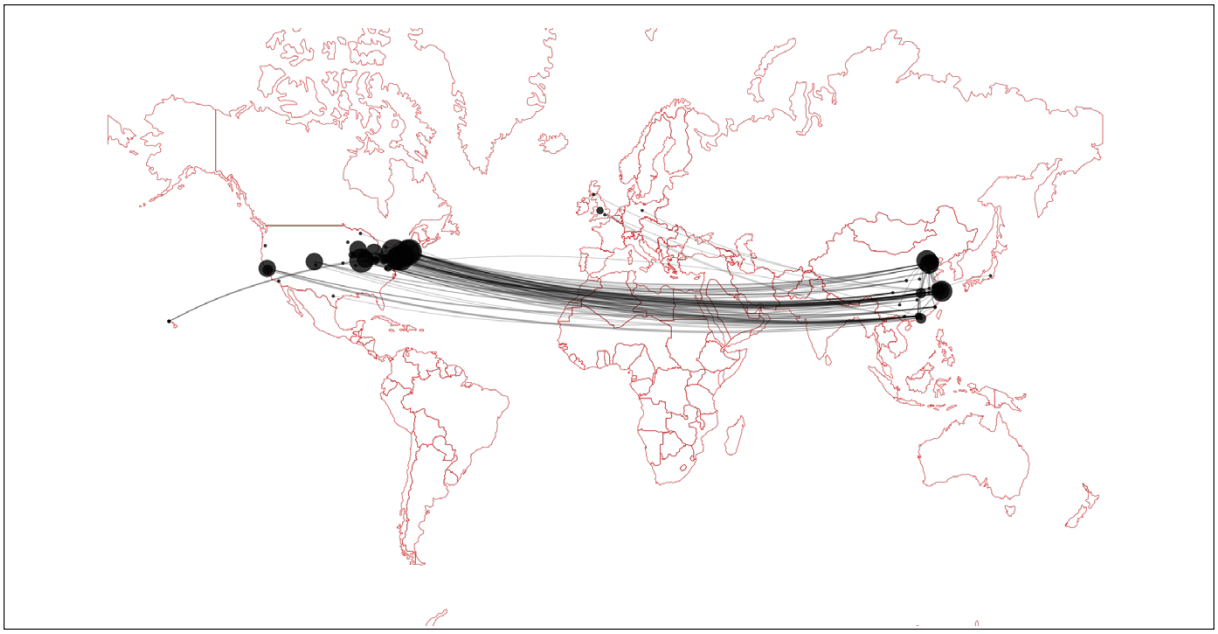


Fig. 5. Visual representation of mobility of foreign-educated students using Palladio

Fig. 6. Comparison of birthplace (left side) and places of employment (right side) using Palladio

The research aspect of mobility can also be analysed when looking at the employments of the engineers and comparing these locations with their native places (see figure 6). Such a visualization can reveal domestic workplace migration among professionals. Apart from the phenomenon of urbanization, it can show old and new core regions of importance to the work of engineering. By querying the data pool for certain dates of employment, such an analysis could further be divided into defined time slices to show core / periphery developments through the decades.

The patent form of visualization for CERD is that of network graphs. Relations between different engineers can be displayed using common memberships in associations, common workplaces, or common alma maters as edges. As the amount of ready data is still small at this stage of the project, no meaningful proof of concept for network visualization can be presented in this working paper. Nonetheless, the analysis of professional networks makes one of the primary potentials of the database in the future.

NOTES

NOTES

Leipzig Research Centre Global Dynamics

Working paper series des SFB 1199 an der Universität Leipzig No. 21

ISBN: 978-3-96023-220-9
ISSN: 2510-4845

Universität Leipzig
SFB 1199

E-Mail: sfb1199@uni-leipzig.de

http://research.uni-leipzig.de/~sfb1199

UNIVERSITÄT LEIPZIG
Leipzig Research Centre Global Dynamics

Leibniz-Institut für Länderkunde

Leibniz-Institut für Geschichte und Kultur des östlichen Europa

TECHNISCHE UNIVERSITÄT DRESDEN

Funded by
DFG Deutsche Forschungsgemeinschaft

Leipzig Research Centre
Global Dynamics
ReCentGlobe