



1913 Transnational Chinese
Business Directory
- a Geo-Linguistic Exploration

A Non-programmer's Experience

Helen Tang, University of Toronto

About the Book

INTERNATIONAL CHINESE BUSINESS DIRECTORY

OF THE WORLD

FOR THE YEAR

1913



A Comprehensive List of Prominent Chinese Firms and Individuals

in parts of

China, Japan, India Proper, Indo China, Straits Settlements, Malay Peninsula, Siam, Java, Sumatra, Borneo, Africa, Australia, New Zealand, Hawaii, Philippines, Canada, United States, Mexico, Central America, South America and West Indies, Etc.

英屬加拿大哥林比亞省
BRITISH COLUMBIA, CANADA.

店名 Name.	生意 Business.	門牌號數街名 No. Street.
-------------	-----------------	-----------------------

很卜碌埠 CRANBFOOK—(Continued)

馬利	雜貨	Mar Lee, General Merchandise.....P. O. Box 842
三利	雜貨	Sam Lee, General Merchandise.....P. O. Box 122
三益	雜貨	Sam Yick & Co., General Merchandise.....P. O. Box 114
東成	雜貨	Tong Sing, General Merchandise.....
悅利	雜貨	Yat Lee, General Merchandise.....P. O. Box 114

雷把倫埠 CUMBERLAND

寶豐	雜貨	Bow Fung, Groceries.....P. O. Box 121
勝記	雜貨	Charlie Sing Kee, General Merchandise.....
蚨源	雜貨	Foo Yuen, General Merchandise and Drugs.....P. O. Box 5
炎記	酒樓	Gum Kee, Restaurant.....
合記	雜貨	Hop Kee, Groceries.....
恒和	雜貨	Hung Wo & Co., Groceries.....P. O. Box 187
其豐	雜貨	Ka Fung & Co., General Merchandise.....P. O. Box 173
廣利	雜貨	Kwong Wah, Dry Goods.....P. O. Box 138
麗豐	豬肉	Lai Fong, Butcher.....
麗源	雜貨	LAI YUEN & CO., General Merchandise.....P. O. Box 137
長老會		Presbyterian Church.....P. O. Box 154
均興	雜貨	Quan Hing, Groceries.....
廣興	雜貨	Quong Hing, Groceries.....P. O. Box 201
廣美隆	油集	QUONG MEE LUNG, Groceries.....P. O. Box 133
成利	雜貨	Sang Lee, General Merchandise.....P. O. Box 98
勝記	雜貨	SING KEE, General Merchandise.....P. O. Box 116
萃芳	餐館	Sueng Fong, Restaurant.....
泰利	豬肉	Tai Lee, Butcher.....
泰安	雜貨	Tai On, General Merchandise.....P. O. Box 117

A Typical
Page from
the Book



The Digital Outcome – a Project Website

URL:

<https://greenclocloud-tang.github.io/chinese-business-directory/>



Purpose of the Project

3. An Archive of Historical Dialect-Based Chinese Transliterations of Geographical Names

孫文大哥黃芸蘇君於七月初二日動程，先往鉢崙、舍路、士卜公頃、抓李抓罇、迫加斯地、杭定頓、南巴、貝士卜提、爹罇、惡頓、梳力、洛士丙令、典化、懇士斯地、聖壘。

馮自由（1939）。〈美洲致公堂與《大同報》〉。載於馮自由，《革命逸史》（頁196–235）。長沙：商務印書館。

Purpose of the Project

鉢崙 = Portland
士卜公頃 = Spokane
舍路 = Seattle
抓李抓罇 = Walla Walla
迫加斯地 = Baker City
杭定頓 = Huntington
南巴 = Nampa
貝士卜提 = Boise Port
懇士斯地 = Kansa City
梳力 = Salt Lake City
典化 = Denver (Colorado)
聖壘 = St. Louis

Project Workflow

Extract Data

- AI assisted OCR & programming (Python)

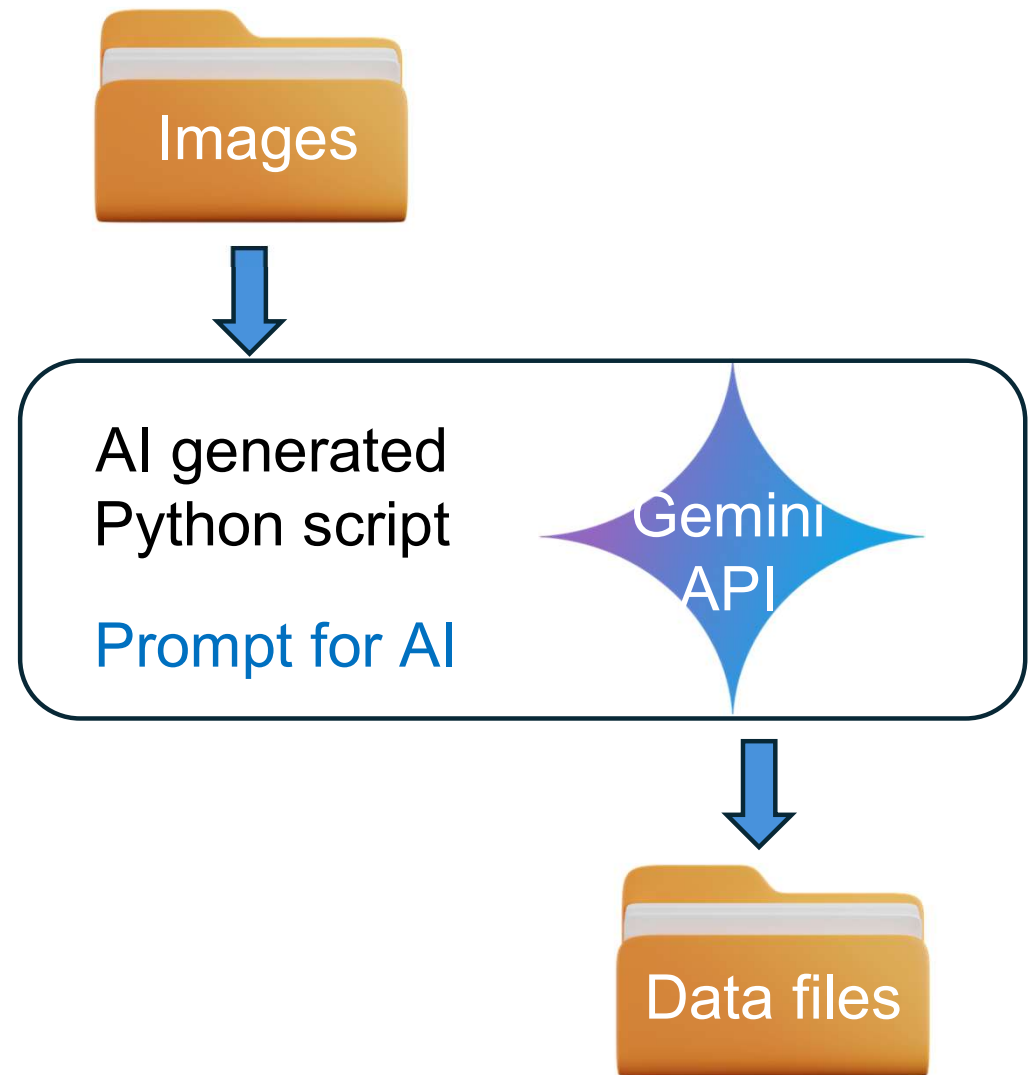
Process Data

- Manual review
- AI assisted Geocoding (Python)

Create Data Table & Build Website

- AI assisted programming (JavaScript, HTML, CSS)
- Static Site Generator
- Github Pages

AI OCR - AI Automated Data Extraction



Prompt for AI

I need to extract the table data from the image below.

Please note the following:

1. It contains a main title, which is a province in Canada in both traditional Chinese and English. It should be populated to two columns "Province Chinese" and "Province English".
2. Each table on the image contains the directory of Chinese business in a town in that province. The town name is also in traditional Chinese and English above each table. They should go to columns "Town Chinese" and "Town English".
3. Each table contains the Chinese name and English name of the business.
4. For the Chinese name of the business, there are two parts with some space in between. The 1st part is the name of the business, the 2nd part is the type of the business. They should go to columns "Business name Chinese" and "Business type Chinese".
5. For the English name of the business, there are also two parts separated by ",". The 1st part is the name of the business, the 2nd part is the type of the business (eg. laundry, groceries). They should go to columns "Business name English" and "Business type English".
6. One more column for the address which often is a P.O Box number.
7. One more column "Page number" which is the page number of the image, and should be extracted from either the upper left corner or upper right corner of the image.
8. If a column is missing data, please put "no data" as a place holder there.
9. Use ";" as the delimiter in the output cvs file.

The header of the cvs should be in this order: Province Chinese; Province English; Town Chinese; Town English; Business name Chinese; Business type Chinese; Business name English; Business type English; Address; Page number



Data Extraction (tabular data)

Province Chinese	Province English	Town Chinese	Town English	Business name Chinese	Business type Chinese	Business name English	Business type English	Address	Page number
英屬加拿大 比亞省	BRITISH COLUMBIA, CANADA	很卜碌埠	CRANBFO OK	馬利	雜貨	Mar Lee	General Merchandise	P. O. Box 842	1356
英屬加拿大 比亞省	BRITISH COLUMBIA, CANADA	很卜碌埠	CRANBFO OK	三利	雜貨	Sam Lee	General Merchandise	P. O. Box 122	1356
英屬加拿大 比亞省	BRITISH COLUMBIA, CANADA	很卜碌埠	CRANBFO OK	三益	雜貨	Sam Yick & Co.	General Merchandise	P. O. Box 114	1356
英屬加拿大 比亞省	BRITISH COLUMBIA, CANADA	很卜碌埠	CRANBFO OK	東成	雜貨	Tong Sing	General Merchandise	no data	1356
英屬加拿大 比亞省	BRITISH COLUMBIA, CANADA	很卜碌埠	CRANBFO OK	悅利	雜貨	Yat Lee	General Merchandise	P. O. Box 114	1356
英屬加拿大 比亞省	BRITISH COLUMBIA, CANADA	叻把崙埠	CUMBERL AND	寶豐	雜貨	Bow Fung	Groceries	P. O. Box 121	1356

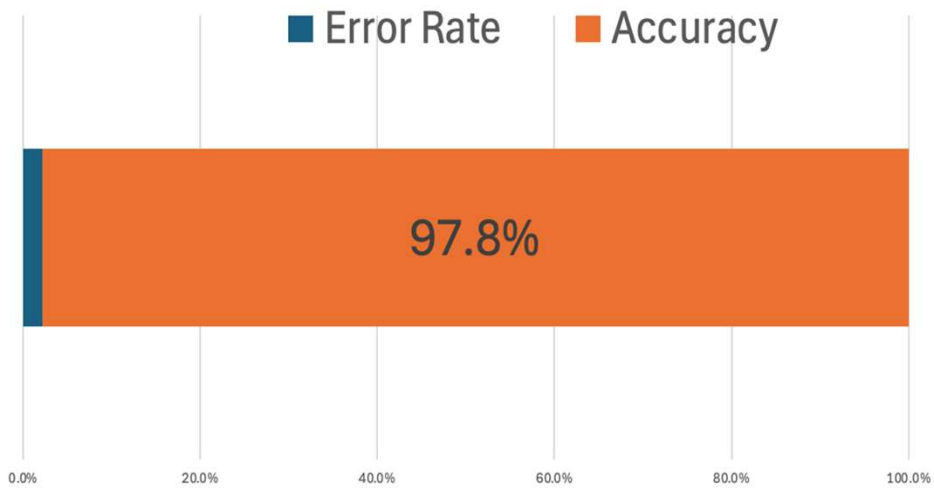


AI OCR - AI Automated Data Extraction

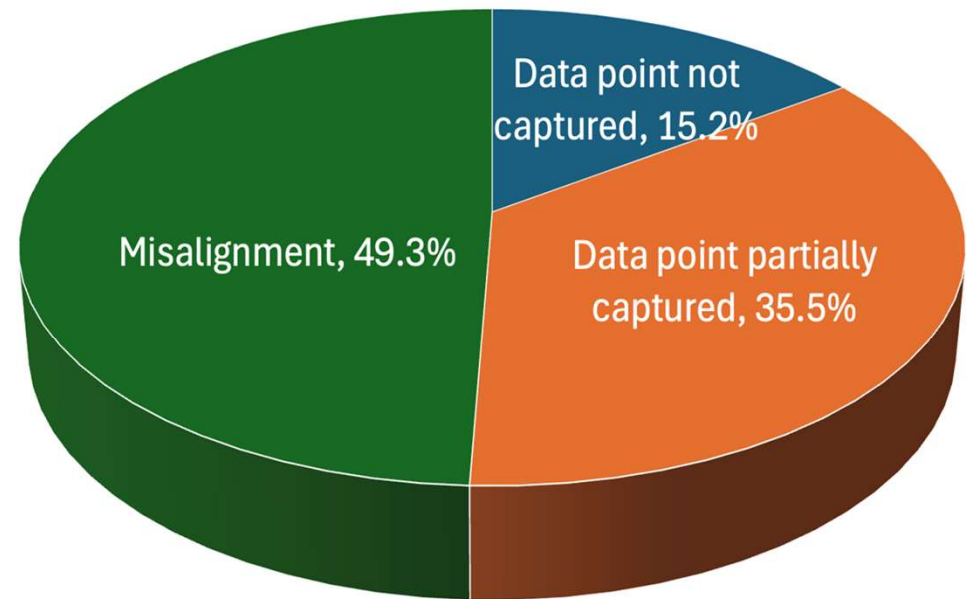
Data Extracted

	Entries (Rows)	Data Points (Cells)
China	28,000	280,000
Outside China	19,000	190,000
Total	47,000	470,000

AI OCR Accuracy



Error Types



AI OCR - AI Automated Data Extraction

To improve the accuracy:

1. High resolution images
2. A well-designed & specific prompt
3. Choosing the right LLMs. Gemini performs surprisingly well to OCR Traditional Chinese characters, even for rare variants (異體字).

罇 囉 畠 釐 贊 罝 懇 齷 孺 鑿 潛

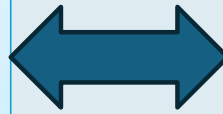
Working with AI – a Partnership

My Role: The Architect (What & Why)

Project Design: broke one big idea into small, clear tasks.

Manual Integration: ran the code and moved the data between steps.

Final Approval: tested the results and provided feedback to debug.



AI's Role: The Builder (How)

Technical Writing: wrote a total of 1,000+ lines of code across four languages.

Instant Troubleshooting: debugged based on feedback.

Rapid Updates: adjusted the design or logic based on feedback.

Findings from the
data

–
Dialect-based
Transliterations

Regional dialects or accents (mostly
Cantonese and Toishanese)

Place Name	Transliteration
------------	-----------------

Alameda:	亞拉蜜打埠, 亞罇必打埠
----------	--------------

Allegheny:	亞利堅釐埠, 爹利堅汝埠
------------	--------------

Place Name	Transliteration	Toishanese	Cantonese
------------	-----------------	------------	-----------

Albuquerque	答部隙埠	答: ap	dap
-------------	------	-------	-----

Anaconda	典地港打埠	典: ein	dein
----------	-------	--------	------

Findings from the data

–

Dialect-based Transliterations

Phonetic accuracy

Place Name	IPA	Transliteration
Quesnel	kwə'nɛl	改市尼埠
Clayoquat	'klækwɒt	企李島括埠

Absence of the “r” and the syllable-final “l” sound

Place Name	Transliteration	Note
Arizona	亞利笏拿省	“ri” = 利 /li/
Colorado	哥羅拉多省	“ra” = 拉 /la/
Barkerville	百架委利埠	“ville” = 委利 /vili/
Lowell	羅委路埠	“well” = 委路 /vilu/

Findings from the
data
—
Dialect-based
Transliterations

Intentional addition or omission of
syllables

Lynn = 阿連埠

Butte = 猫失地埠 (失地=City)

Hueneme 懷念埠

Philadelphia = 費城埠

Los Angeles = 羅生埠

Guadalupe = 和老比埠

Thank you!

Helen Tang

helent.tang@utoronto.ca

Project Website:

<https://greenclocloud-tang.github.io/chinese-business-directory/contact/>

