Improving Accessibility and Understandability of Open Research Data
-- Metadata and Data Guide to the CCVG Data Project at the University of Pittsburgh Library System

Fanghui Xiao, School of Computing and Information
Haihui Zhang, University of Pittsburgh Library System

University of Pittsburgh
Contemporary Chinese Village Gazetteer Data Project (CCVG Data)

- Led by East Asian Library (EAL), University Library System (ULS)
  - Supported by the Provost in 2018
  - Collaborate with School of Computing and Information
- Aim to extract data from ULS’s Chinese village gazetteer collection
  - 12 Topics include Natural Information, Population, Politics and Management, Economics, Education, etc.
Project Overview

Focus on administrative villages in China
Present: 1,800 villages
Cover 31 provinces (including municipalities), 212 cities, 510 counties
Goal: reach 2,500-3,000

Extracted Data in 18 categories and 147 subcategories
E.g., Number of Civil Mediations
E.g., Agricultural / Industrial / Household / Commercial Electricity Price
# Datasets

18 datasets in total

<table>
<thead>
<tr>
<th>CSV (Gazetteer Information 村志信息)</th>
<th>CSV (Ethnic Groups - Range 民族 - 年代段)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>Data</td>
</tr>
<tr>
<td>Available under License Creative Commons Attribution Non-commercial Share Download (117kB)</td>
<td>Available under License Creative Commons Attribution Non-commercial Share Alike Download (2kB)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CSV (Village Information 村庄信息)</th>
<th>CSV (Ethnic Groups - Yearly 民族 - 年度)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>Data</td>
</tr>
<tr>
<td>Available under License Creative Commons Attribution Non-commercial Share Download (460kB)</td>
<td>Available under License Creative Commons Attribution Non-commercial Share Alike Download (925kB)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CSV (Natural Environment 自然环境)</th>
<th>CSV (Population and Migration - Range 人口与人口迁移 - 年代段)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>Data</td>
</tr>
<tr>
<td>Available under License Creative Commons Attribution Non-commercial Share Download (86kB)</td>
<td>Available under License Creative Commons Attribution Non-commercial Share Alike Download (91kB)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CSV (Natural Disasters 自然灾害)</th>
<th>CSV (Population and Migration - Yearly 人口与人口迁移 - 年度)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>Data</td>
</tr>
<tr>
<td>Available under License Creative Commons Attribution Non-commercial Share Download (463kB)</td>
<td>Available under License Creative Commons Attribution Non-commercial Share Alike Download (3MB)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CSV (Last Names 姓氏)</th>
<th>CSV (Military, Politics and Management - Range 军事, 政治, 管理 - 年代段)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>Data</td>
</tr>
<tr>
<td>Available under License Creative Commons Attribution Non-commercial Share Download (95kB)</td>
<td>Available under License Creative Commons Attribution Non-commercial Share Alike Download (246kB)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CSV (Year of First Availability or Purchase 首次拥有年份)</th>
<th>CSV (Military, Politics and Management - Yearly 军事, 政治, 管理 - 年度)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>Data</td>
</tr>
<tr>
<td>Available under License Creative Commons Attribution Non-commercial Share Download (83kB)</td>
<td>Available under License Creative Commons Attribution Non-commercial Share Alike Download (1MB)</td>
</tr>
</tbody>
</table>
## Project Overview — Data

### Example: CSV (Economy - Yearly 经济 - 年度)

<table>
<thead>
<tr>
<th>Village Code</th>
<th>Category Description</th>
<th>Division 1</th>
<th>Subdivision</th>
<th>Unit</th>
<th>1996</th>
<th>1997</th>
<th>1998</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>村志书名</td>
<td>集体经济收入 Collective Econo. 总 Total</td>
<td>元 yuan</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>村志书名</td>
<td>耕地面积 Cultivated Area</td>
<td>亩 mu</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>1539</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>村志书名</td>
<td>人均收入 Per Capita Income</td>
<td>元 yuan</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>村志书名</td>
<td>人均居住面积 Per Capita Living Space</td>
<td>平方米 square n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>86</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>村志书名</td>
<td>集体经济收入 Collective Econo. 种植业 Agriculture</td>
<td>元 yuan</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>村志书名</td>
<td>集体经济收入 Collective Econo. 林业 Forestry</td>
<td>元 yuan</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>村志书名</td>
<td>集体经济收入 Collective Econo. 牧业 Animal Husbandry</td>
<td>元 yuan</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>村志书名</td>
<td>集体经济收入 Collective Econo. 副业 Miscellaneous</td>
<td>元 yuan</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>村志书名</td>
<td>音西村志</td>
<td>人均收入 Per Capita Income</td>
<td>元 yuan</td>
<td>4545</td>
<td>4920</td>
<td>5135</td>
<td>5585</td>
<td>6648</td>
<td>6500</td>
<td>6525</td>
<td>6525</td>
</tr>
<tr>
<td>村志书名</td>
<td>总产值 Gross Output Value 总 Total</td>
<td>万元 10K yuan</td>
<td>27010</td>
<td>36490</td>
<td>38200</td>
<td>40200</td>
<td>48000</td>
<td>48600</td>
<td>48850</td>
<td>48850</td>
<td>48850</td>
</tr>
<tr>
<td>村志书名</td>
<td>耕地面积 Cultivated Area</td>
<td>亩 mu</td>
<td>2527</td>
<td>2385</td>
<td>2307</td>
<td>2128</td>
<td>2110</td>
<td>1628</td>
<td>1625</td>
<td>1625</td>
<td>1625</td>
</tr>
<tr>
<td>村志书名</td>
<td>人均收入 Per Capita Income</td>
<td>元 yuan</td>
<td>7174</td>
<td>7000</td>
<td>6503</td>
<td>6658</td>
<td>7060</td>
<td>7066</td>
<td>7650</td>
<td>7650</td>
<td>7650</td>
</tr>
<tr>
<td>村志书名</td>
<td>郭家沟村志</td>
<td>总产值 Gross Output Value 总 Total</td>
<td>万元 10K yuan</td>
<td>13042</td>
<td>10485</td>
<td>12178</td>
<td>15336</td>
<td>17900</td>
<td>18120</td>
<td>16698</td>
<td>16698</td>
</tr>
<tr>
<td>村志书名</td>
<td>人均收入 Per Capita Income</td>
<td>元 yuan</td>
<td>6790</td>
<td>6874</td>
<td>7018</td>
<td>7380</td>
<td>7405</td>
<td>7508</td>
<td>7800</td>
<td>7800</td>
<td>7800</td>
</tr>
<tr>
<td>村志书名</td>
<td>郭家沟村志</td>
<td>总产值 Gross Output Value 种植业 Agriculture</td>
<td>万元 10K yuan</td>
<td>151</td>
<td>156</td>
<td>280</td>
<td>350</td>
<td>300</td>
<td>640</td>
<td>655</td>
<td>655</td>
</tr>
<tr>
<td>村志书名</td>
<td>人均收入 Per Capita Income</td>
<td>元 yuan</td>
<td>7174</td>
<td>7000</td>
<td>6503</td>
<td>6658</td>
<td>7060</td>
<td>7066</td>
<td>7650</td>
<td>7650</td>
<td>7650</td>
</tr>
<tr>
<td>村志书名</td>
<td>郭家沟村志</td>
<td>总产值 Gross Output Value 种植业 Agriculture</td>
<td>万元 10K yuan</td>
<td>n/a</td>
<td>n/a</td>
<td>280</td>
<td>496</td>
<td>451</td>
<td>772</td>
<td>955</td>
<td>955</td>
</tr>
<tr>
<td>村志书名</td>
<td>人均收入 Per Capita Income</td>
<td>元 yuan</td>
<td>6790</td>
<td>6874</td>
<td>7018</td>
<td>7380</td>
<td>7405</td>
<td>7508</td>
<td>7800</td>
<td>7800</td>
<td>7800</td>
</tr>
<tr>
<td>村志书名</td>
<td>郭家沟村志</td>
<td>总产值 Gross Output Value 牧业 Animal Husbandry</td>
<td>万元 10K yuan</td>
<td>n/a</td>
<td>n/a</td>
<td>150</td>
<td>72</td>
<td>69</td>
<td>576</td>
<td>250</td>
<td>250</td>
</tr>
<tr>
<td>村志书名</td>
<td>人均收入 Per Capita Income</td>
<td>元 yuan</td>
<td>6790</td>
<td>6874</td>
<td>7018</td>
<td>7380</td>
<td>7405</td>
<td>7508</td>
<td>7800</td>
<td>7800</td>
<td>7800</td>
</tr>
<tr>
<td>村志书名</td>
<td>郭家沟村志</td>
<td>总产值 Gross Output Value 渔业 Fishery</td>
<td>万元 10K yuan</td>
<td>n/a</td>
<td>n/a</td>
<td>80</td>
<td>151</td>
<td>132</td>
<td>87</td>
<td>102</td>
<td>102</td>
</tr>
<tr>
<td>村志书名</td>
<td>人均收入 Per Capita Income</td>
<td>元 yuan</td>
<td>6790</td>
<td>6874</td>
<td>7018</td>
<td>7380</td>
<td>7405</td>
<td>7508</td>
<td>7800</td>
<td>7800</td>
<td>7800</td>
</tr>
</tbody>
</table>
# Data Dictionary

## Contents

### I. General Introduction

- A. Data Selection and Collection ........................................ 8
- B. Data Normalization .................................................. 9
- C. Gazetteer Key ....................................................... 10

### II. Introduction to Individual Datasets ................................ 10

- A. 村志信息 Gazetteer Publication Information .......................... 10
  - Bib ID .......................................................... 10
  - 书名 - 汉字 Gazetteer Title - Chinese Characters .................. 10
  - 书名 - 汉语拼音 Gazetteer Title - Hanyu Pinyin .................. 10
  - 出版类型 Publication Type ................................... 10
  - 年代 Year of Publication ...................................... 11
- B. 村庄信息 Village Information ..................................... 11
  - 村代码 Village Code ........................................... 11
  - 村名 - 汉字 Village Name - Chinese Characters ................. 11
  - 村名 - 汉语拼音 Village Name - Hanyu Pinyin ................. 11
  - 省 - 汉字 Province - Chinese Characters ........................ 11
  - 省 - 汉语拼音 Province - Hanyu Pinyin ........................ 12
  - 市 - 汉字 City - Chinese Characters ............................ 12
  - 市 - 汉语拼音 City - Hanyu Pinyin ............................. 12
  - 县 / 区 - 汉字 County / District - Chinese Characters .......... 12
  - 县 / 区 - 汉语拼音 County / District - Hanyu Pinyin .......... 12
  - 村庄总面积 单位 Total Area Unit ................................ 12
  - 村庄总面积 Total Area ....................................... 13
  - 经纬度 - 格式 Latitude and Longitude - Format ................ 13
  - 纬度 Latitude .................................................. 13
  - 经度 Longitude ................................................ 13
  - 距附属县城距离单位 Distance to Affiliated to the county town - Unit ........................................................................ 13

Data Dictionary— General Introduction

Data Selection and Collection

E.g., All data are taken from the village gazetteers, except administrative levels and village code, which from related government website.
E.g., Data are recorded using the original measurement units provided in the gazetteer.

Data Normalization

- Data Selection and Collection
  - E.g., All data are taken from the village gazetteers, except administrative levels and village code, which from related government website.
  - E.g., Data are recorded using the original measurement units provided in the gazetteer.

Gazetteer Key

1. All datasets include:
   a. Gazetteer Code
      - Gazetteer Code: Four Arabic numerals. Codes are sequential beginning at 0001 and are assigned to the gazetteers at random. The codes are used as a key during data entry.
   b. Gazetteer Title
      - Gazetteer Title (Chinese characters): Taken from the title page of the gazetteer.
Data Dictionary—Introduction to Individual Datasets

Example: 村志信息 Gazetteer Publication Information

A. 村志信息 Gazetteer Publication Information

Bib ID
为匹兹堡大学图书馆编目记录号（ULS record ID）
University Library System, University of Pittsburgh record ID.

书名 - 汉字 Gazetteer Title - Chinese Characters
取自村志书名页。
Taken from the title page of the gazetteer.

出版类型 Publication Type
包括汉字及汉语拼音。汉语拼音字段依美国国家图书馆主题词（LC classification, subject heading）为准。

出版年 Year of Publication
正式出版物的出版年取自版权页；非正式出版及内部出版物的出版年则参考书内信息提取。
For “formal” publications, the publication year is selected from the copyright page. For “informal” and “internal” publications, the publication year is selected from publication information given elsewhere in the gazetteer, such as in the Preface or Explanatory Notes. For more information on the distinction between “formal,” “informal,” and “internal,” see below.
Example: 经济 Economy

M. 经济 Economy (Range)

- 总产值 Gross Output Value
- 第二产业 Second Industry
- 集体经济收入 Collective Economic Income
- 耕地面积 Cultivated Area
- 粮食总产量 Total Grain Output
- 用电量 Electricity Consumption
- 电价 Electricity Price
- 用水量 Water Consumption
- 水价 Water Price
- 人均收入 Per Capita Income
- 人均居住面积 Per Capita Living Space
- Start Year
- End Year
- Data

Units

- 百分比 (%) percentage
- 元 yuan
- 百元 100 yuan
- 万元 10K yuan
- 千万元 10 millions
- 亿元 100 millions

第一产业 Primary Industry
Includes agriculture, forestry, animal husbandry, and fisheries.

种植业 Agriculture
Indicates crop farming.
- 蔬菜 Vegetables
- 水果 Fruit
- 粮食 Grain

林业 Forestry

牧业 Animal Husbandry

副业 Miscellaneous
- 仓储业 Warehouse
Database Metadata
Database Metadata

- Descriptive Metadata
  - Description
  - Contributor
  - Publication date

- Technical Metadata
  - Creation date
  - Last updated date
  - Data format

- Preservation Metadata
  - Identifier
  - Language
  - File type

- Rights Metadata
  - Copyright status
  - License terms
  - Rights holder

Metadata Attributes
Data Guide

Survey sample: 33 participants

How important is the following contextual information in the current CCVDG data dictionary in helping you reducing confusion, uncertainty, and ambiguity when working with datasets?

Data collection and creation process

Information about publisher

Citation information
How important is the following information that might help you gain insights when working with datasets?

Publisher's purposes for collecting and publishing the datasets

Potential uses of a dataset

Examples of how the data has been used
How important is the following information that might help you find connections between datasets?

The relationships (scenario-based connections) with other datasets within the same data repository

Cross-linking to other sites' data that related to the data that you are looking at
Data Guide

How important is the following information that might help you **structure the datasets**? (knowing how data inside the current database is organized?)

- **Database schema**
  - Extremely important: 9.1%
  - Very important: 50.0%
  - Moderately important: 36.4%
  - Slightly important: 0.0%
  - Not at all important: 4.5%

- **History of data formats**
  - Extremely important: 10.0%
  - Very important: 35.0%
  - Moderately important: 40.0%
  - Slightly important: 10.0%
  - Not at all important: 5.0%
Data Guide

Overall, to what extent do you think a Data Guide can help with using the datasets?

- Data collection and creation process
- Information about publisher
- Citation information
- Publisher's purposes for collecting and publishing the datasets
- Potential uses of a dataset
- Examples of how the data has been used
- The relationships (scenario-based connections) with other datasets within the same data portal
- Cross-linking to other sites' data that related to the data that you are looking at
- Database schema
- History of data format
Thank you & Questions?