University of Pittsburgh

可作 可为 可行 – CCVD数据库创建及其图书馆员的作用

匹兹堡大学东亚图书馆

张海惠

2019/8

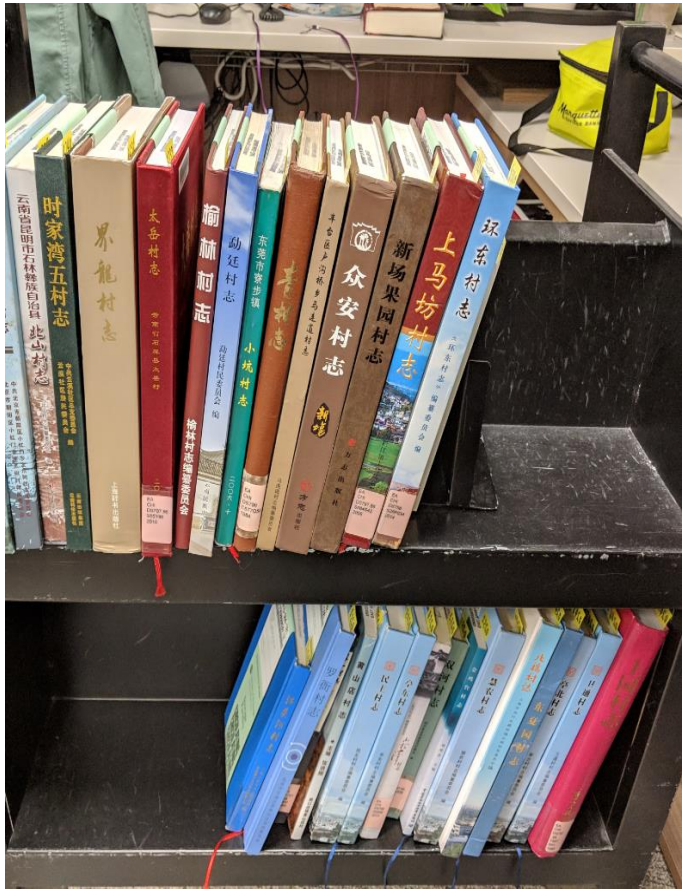
PITT数字人文项目回顾

2002 – 2013 初试、困顿
2014 – 2017 实践
2018 – 探索、创新

- **文本数字化转换(Digitizing and transforming)**
  - Modern China Studies 现代中国研究 2002
  - Szeming Sze Papers 施思明文稿 2014
  - Chinese Land Records 地契文献 2015
  - Chinese Political Prisoners 民国政治犯照片 2016
  - 2017 Oversea Chinese Student Newsletters 海外中国留学生校刊 2017

- **原生数字化项目(Born digital initiative)**
  - Oral history口述史 2015-
  - Contemporary Chinese Village Data 数字村庄 2018 -

关于村志



- 村村镇市行理产史和志志是自是县单历人产术地然地三位史物，价方村方级为、等有值，志为志志记经方着的记书书述济面特具一述的的对、的殊有种，围要伸，俗、况历他就的组和全、史书是志成补面文是价籍以书部充。盘化、十值、不某。分。点、分、可一村，它乡教珍文替政乡、层物历值功行和省基地、的价的个志是以村育贵化代

- 引自“中国村志网”
- http://www.cunzhiwang.com/page/guanyucunzhi.html

数字村庄

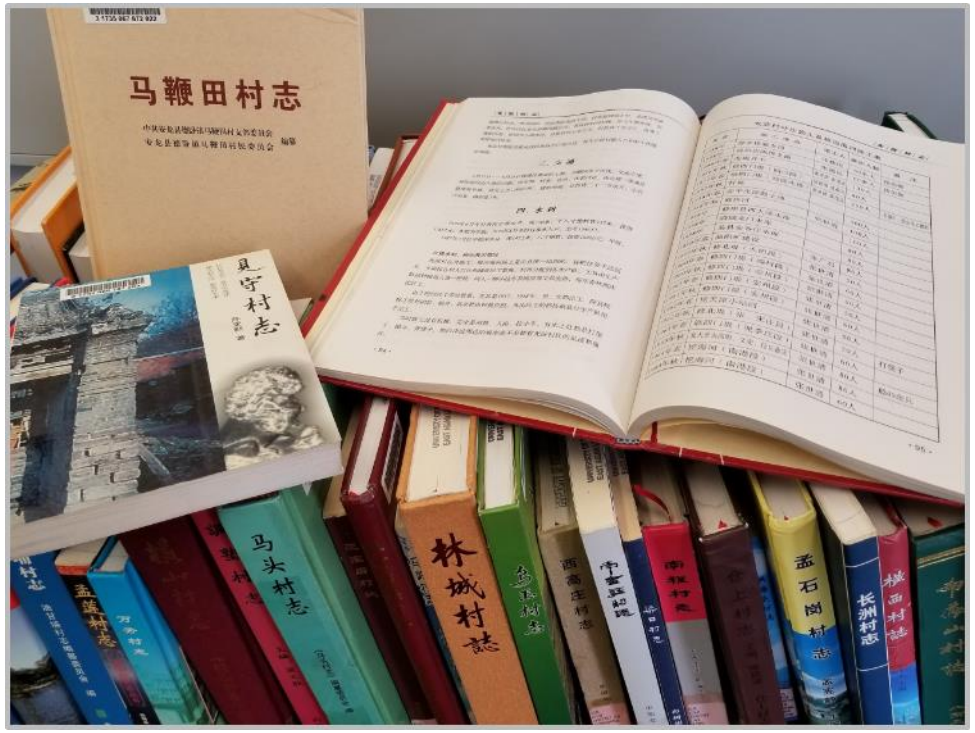
Contemporary Chinese Village Database (CCVD)

- **特性**
  - 原生数字人文项目
  - 自主设计开发
  - Open Access
  - 可持续发展

- **目标**
  - 随机选择2,000 – 2,500村庄
  - 提取原始数据
  - 提供检索和下载数据
  - 2021完成

数据库创建目的

- 中国村庄建设数据可视化
- 为教学与科研提供数据支持
- 为政策制定者和投资者提供数据参考
- 开拓图书馆数字人文的新思路
- 探讨图书馆员新的工作内容和作用



提取数据类别

村庄信息

- 经纬度、面积、到县城的距离。。。。。。

历年自然灾害

- 风灾、旱灾、水灾、地震、沙尘暴、涝灾。。。。。。

首次拥有年

- 供电、自来水、有线广播、电话、电视机。。。

民族

- 户、人、百分比

姓氏

- 姓氏总数、前五大姓

提取数据类别2

人口

- 户数、男、女、总人口、迁入、迁出、流动人口。。。

计划生育

- 出生、死亡、自然增长率、计划生育率。。。

政治与管理

- （土改）阶级成分、民事调解。。。

经济

- 总产值、经济总收入、耕地面积、粮食产量、人均收入和居住面积。。。

教育

- 小学/中学/高中生在校生、新入学大学生、教师人数、受教育程度。。。

数据类别例1

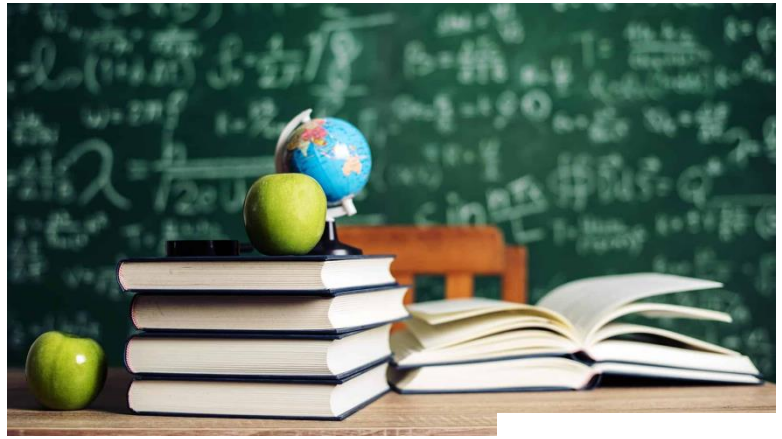
• 经济发展

总产值	总
	第一产业
	农业 (蔬菜/水果/粮食)
	林业
	牧业
	副业 (仓储/运输/建筑)
	渔业
	工业
	第三产业 (商饮/服务业)
集体经济收入	总
	第一产业
	农业 (蔬菜/水果/粮食)
	林业
	牧业
	副业 (仓储/运输/建筑)
	渔业
	工业
	第三产业 (商饮/服务业)
耕地面积	
粮食总产量	

用电量	General
	农业
	工业
	生活 (人/户/村)
	商业
电价	General
	农业
	工业
	生活
	商业
用水量	General
	农业
	工业
	生活
	商业
水价	General
	农业
	工业
	生活
	商业
人均收入	
人均居住面积	

数据类别例2

• 教育



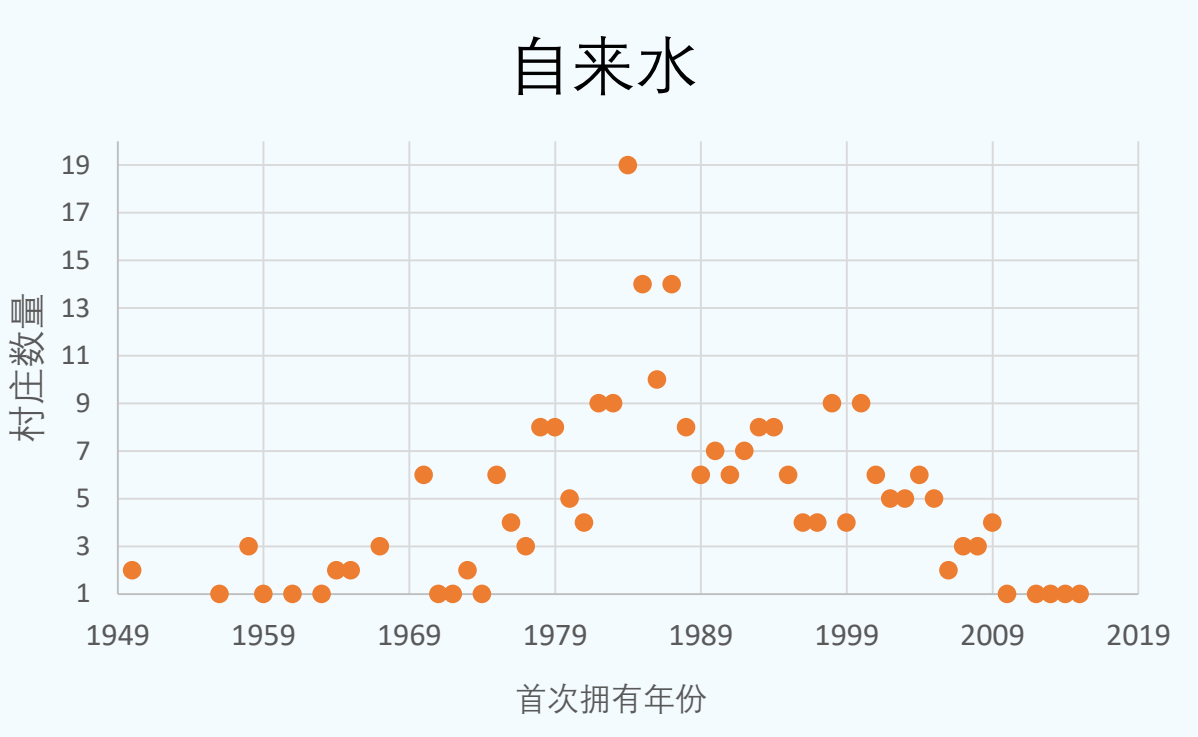
- 在校生
  - 小学
  - 初中
  - 高中
- 新入学生（大学）
- 老师
  - 小学
  - 初中
  - 高中

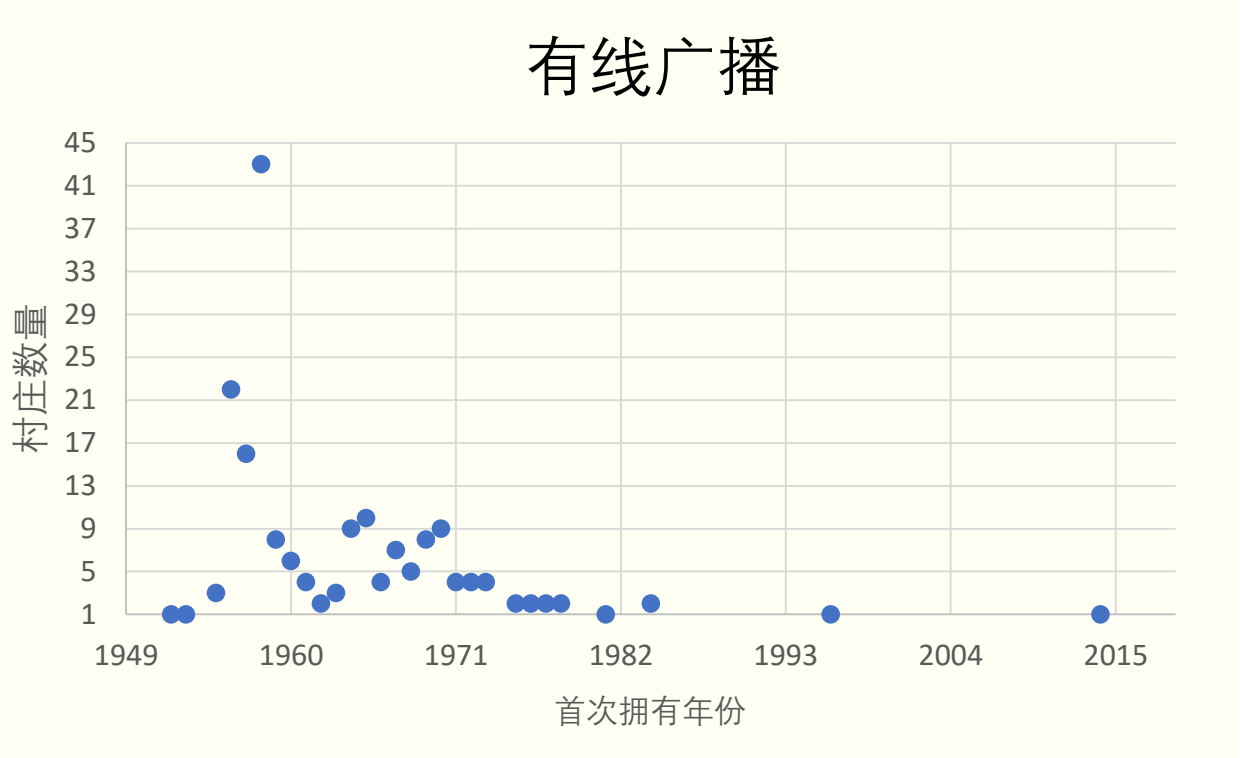
受教育程度

- 文盲（人数/%）
- 小学（人数/%）
- 初中（人数/%）
- 中专高中（人数/%）
- 大专以上（人数/%）

数据应用例1

- 首次拥有年
  - 液化气 • 管道燃气 • 天然气 • 自来水
  - 供电 • 电视机 • 电话机 • 有线广播





数据应用例2

• 姓氏

姓氏总数最多的十个村庄			
村名	省份/直辖市	姓氏总数	前五大姓氏
黄土岗村	北京市	295	王, 卢, 殷, 张, 亢
永联村	江苏省	222	陈, 张, 黄, 王, 陆
新维村	广东省	205	n/a
义东沟村	山西省	201	张, 王, 李, 杨, 赵
南胜村	上海市	201	张, 翁, 夏, 王, 蔡
云溪村	云南省	189	李, 杨, 张, 王, 罗
新房村	陕西省	188	李, 仲, 陈, 王, 张
清河村	山东省	178	n/a
九星村	上海市	170	吴, 赵, 阮, 李, 沈
水师营村	辽宁省	163	王, 刘, 张, 李, 韩

姓氏总数最少的十个村庄			
村名	省份/直辖市	姓氏总数	前五大姓氏
大张营村	河南省	1	张, n/a, n/a, n/a, n/a
马溪村	广东省	4	n/a
魁胆村	贵州省	4	王, 龙, 彭, 陆, n/a
传桂村	海南省	5	洪, 吴, 刘, 汤, 黄
梁口村	河北省	5	郑, n/a, n/a, n/a, n/a
叶坪村	江西省	6	谢, 钟, 张, 刘, 程
堡上村	河南省	7	刘, 郭, 庞, 王, 孙
高山村	广西壮族自治区	7	n/a
洪塘头村	福建省	7	n/a
枣牌刘村	山东省	8	范, 张, 栾, n/a, n/a
五河苗族村	贵州省	8	n/a
莲塘村	福建省	8	n/a

数据应用例3

- 工农业总收入

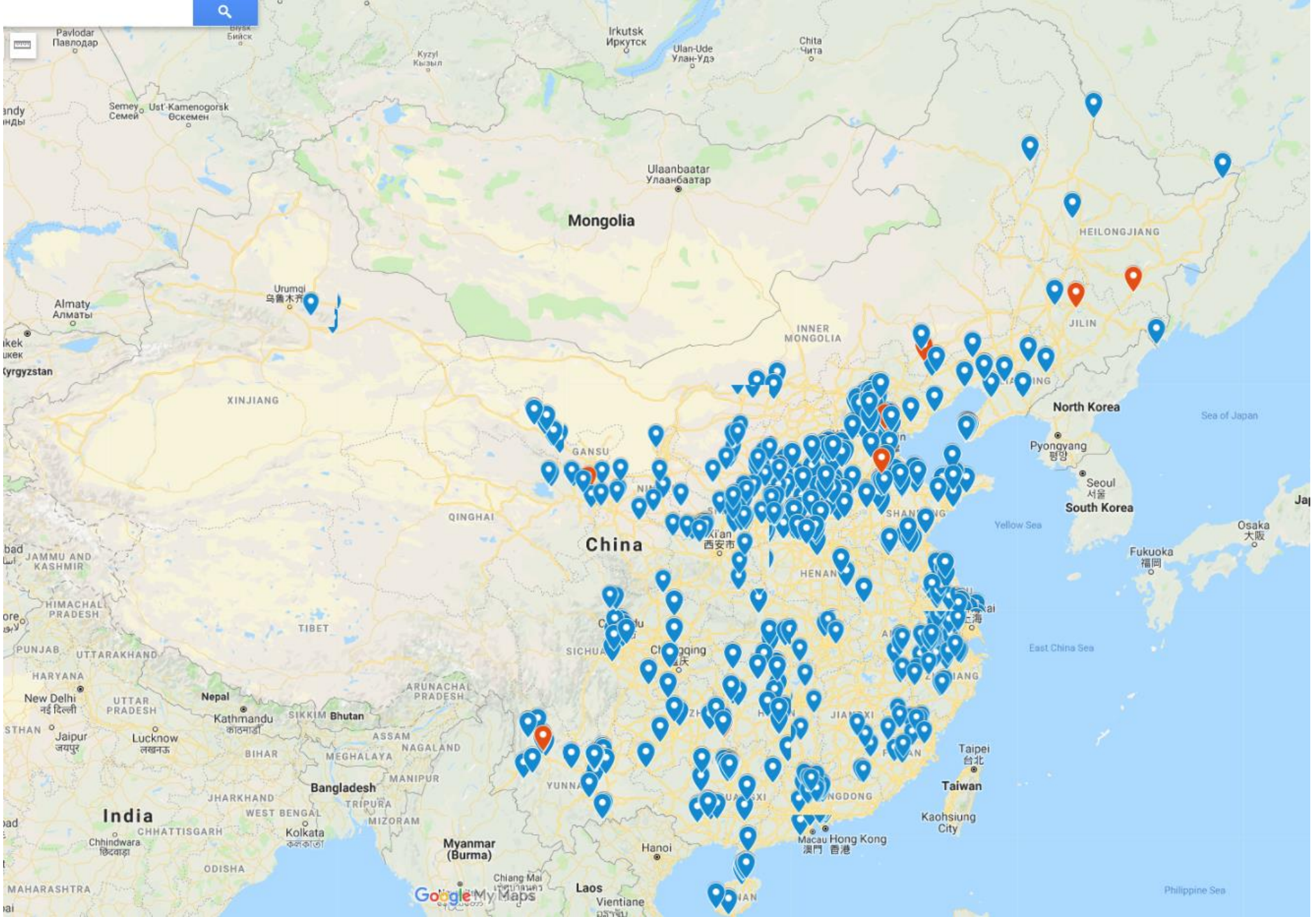
工作流程

- 任意选择村志 Randomly select gazetteers
- 人工提取有效数据 Identify data
- 填写数据提取表格 Fill checklist
- 数据输入专属平台 Data entering
- 数据提取及输入的二次检查 Review
- 下载数据集 Dataset downloading
- 数据格式后期调整和处理 Post-process

项目进展

- 500 村庄数据完成
- 遍布30个省/自治区/直辖市
- 11万个有效数值的数据
- 平均每个村庄220个数据
- 专属网页即将建成

覆盖范围


Mongolia
China
North Korea
South Korea
Taiwan
India
Bangladesh
Myanmar (Burma)
Laos
Nepal
Bhutan
Pyongyang
Seoul
Fukuoka
Osaka
Taipei
Kaohsiung City
Hong Kong
Macau
Hanoi
Vientiane
Chiang Mai
Kathmandu
New Delhi
Lucknow
Jaipur
Kolkata
Chhindwara
Almaty
Urumqi
Irkutsk
Ulaanbaatar
Ulan-Ude
Chita
Kyzyl
Semey
Ust'-Kamenogorsk
Pavlodar
Kyrgyzstan
XINJIANG
TIBET
QINGHAI
GANSU
INNER MONGOLIA
HEILONGJIANG
JILIN
HENAN
SICHUAN
YUNNAN
GUANGDONG
JIANGXI
HIMACHAL PRADESH
JAMMU AND KASHMIR
PUNJAB
UTTARAKHAND
HARYANA
UTTAR PRADESH
BIHAR
JHARKHAND
WEST BENGAL
ODISHA
CHHATTISGARH
MAHARASHTRA
SIKKIM
ASSAM
MEGHALAYA
MANIPUR
TRIPURA
MIZORAM
NAGALAND
ARUNACHAL PRADESH
Sea of Japan
Yellow Sea
East China Sea
Philippine Sea
Google My Maps

CCVD 与图书馆传统数字人文项目

- **文本数字化转换**(Digitizing and transforming)

  - 拥有一定规模的特殊文献馆藏
  - 版权许可
  - 数字化转换并生成纸质文献之外的另一种形式
  - 馆内多部门之间合作

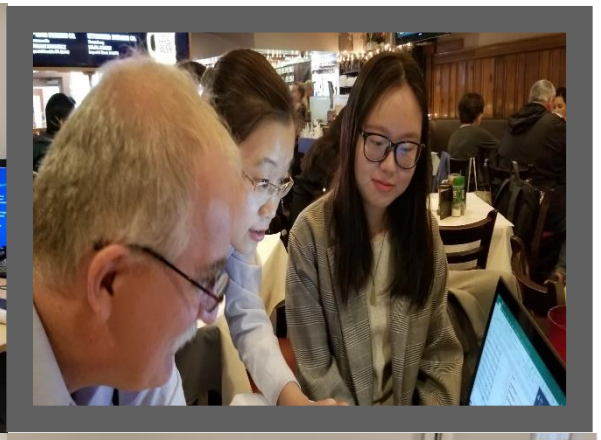
- **原生数字化**(Born digital initiative)

  - 原生、独特的数字人文产品
  - 纸本文献内容的深度揭示
  - 无纸质替代品
  - 版权持有者
  - 馆内多部门及馆外多机构之间合作（如学术顾问委员会、数据专家）

可作　可为　可行

---- 图书馆和图书馆员的角色

- 项目立意者
- 项目发起和执行者
- 各部门和机构协调者
- 产品平台和维护者
- 产品使用指导者
- 产品版权拥有者



开发原生数字人文产品面临的挑战

- **拥有相当规模的特殊馆藏**

- **选题**
  - 特殊馆藏的文献内容（特点和属性）
  - 现有数字人文产品及其空白
  - 教学和研究需求
  - 潜在有数据需求的使用者

- **可行性论证**
  - 产品价值
  - 技术难度
  - 资金

- **具备各类知识储备的团队**
  - 馆内各部门
  - 跨行业专家顾问


University of Pittsburgh
THANKS
谢谢
haihuiz@pitt.edu