Abstract—This paper highlights research on constructing a big archival data called Taiwan Indigenous Peoples Open Research Data (TIPD, see https://osf.io/e4rvz/) based on contemporary census and household registration data sets in 2013-2017 (see http://TIPD.sinica.edu.tw). TIPD utilizes record linkage, geocoding, and high-performance in-memory computing technology to construct various dimensions of Taiwan Indigenous Peoples (TIPs) demographics and developments. Embedded in collecting, cleaning, cleansing, processing, exploring, and enriching individual digital records are archival computational science and data science. TIPD consists of three categories of archival open data: (1) categorical data, (2) household structure and characteristics data, and (3) population dynamics data, including cross-sectional time-series categorical data, longitudinally linked population dynamics data, life tables, household statistics, micro genealogy data, marriage practice and ethnic identity data, internal migration data, geocoded data, etc. TIPD big archival data not only help unveil contemporary TIPs demographics and various developments, but also help overcome research barriers and unleash creativity for TIPs studies.

Keywords—identity; genealogy; in-memory computing; open data; record linkage; TIPD

I. INTRODUCTION

Demand for computational archival science has grown in order to overcome challenges in processing, extracting, and enriching information embedded in rapidly growing big archival data sets [1] [2] [3]. Using archival data to study indigenous peoples is a good example. Although indigenous peoples are small in population size, their crucial role of in different settings of societies and states around the world has long been recognized [4]. Indigenous peoples are recognized as the preservers of global human diversity in the context of ecology, biology, traditional knowledge, culture, and languages. One crucial barrier that still deeply hampers efforts towards the preservation and revival of indigenous peoples’ diversity is the lack of state-of-the-art in-depth data on indigenous peoples.

Taiwan Indigenous Peoples (TIPs) are a branch of Polynesian-Malaysian (or Austronesian) ethnic groups in a genetic and linguistic context. Since the early 17th century, TIPs played a crucial role during the Great Marine Times of East Asia trade. There was a rich body of ethnomorphic, official and academic records on TIPs before 1940. However, the period of 1940-2000 was a data “Dark Ages” for TIPs due to the 1941-45 Pacific War, followed by the 1946-1990 authoritarian rule out of fear of communism and communist infiltration. Persistent lack of TIPs data led TIPs to become isolated, marginalized and thus underdeveloped. Taiwan resumed a TIPs population census in 2000 and began recording TIPs individual records in the household registration system in 2003. This paper highlights the progress and efforts of Taiwan academicians struggling to construct big archival data on contemporary TIPs data by integrating the micro data of the 2000 Taiwan population census and the household registration system.

The research is conducted on the basis of a four-year Joint Research Agreement between Academia Sinica and the Council of Indigenous Peoples in 2013-2016. One important aim of the research is to construct big anonymous Taiwan Indigenous Peoples open research data based on contemporary census and household registration data sets (TIPD, see http://TIPD.sinica.edu.tw; for the TIPD archival big data repository, see https://osf.io/e4rvz/). Serving as the PI, the author demonstrates how data science, i.e., integration of hacking skills, archival computational science, advanced math and statistics knowledge, and domain knowledge expertise, are applied to construct TIPD big data based on Taiwan Household Registration administrative micro data. The paper highlights the theoretical foundation, implementation process, challenges, and ways to overcome research barriers in processing and enriching digital archival big data. Methods embedded in the research include household geocoding, deterministic and probabilistic record linkage, high-performance computing (HPC) facilities, and in-memory computing methods.

II. THEORETICAL FOUNDATION AND COMPUTATIONAL INFRASTRUCTURE

A. Sources of Archival Data

There are various challenges in coping with big data [5]. The main source of data utilized to construct TIPD is Taiwan Household Registration (THR) administrative micro data. THR administrative data are collected for vital official statistics that comprise (1) information on an
individual’s name, personal identification number, birth, death, marriage, education, age, birth place, migration and residential mobility, employment, relationship to household head, spouse name, father name, mother name, ethnicity, etc. and (2) household information like household head name, administrative area codes, address, size, etc. at an aggregate level.

B. Theoretical Foundation

Based on computational archival science and data science, TIPD utilizes record linkage [6], geocoding, and high-performance in-memory computing technology to construct various dimensions of TIPs demographics and developments. Embedded in collecting, cleaning, cleansing, processing, exploring, and enriching individual digital records is data science. Data science is by no means a new field. Rather, it is multi-disciplinary in essence and consists of three necessary components: (1) hacking skills, (2) advanced mathematical and statistical knowledge and skills, and (3) domain knowledge expertise [7].

The practices of the research are based on real world domain knowledge expertise that enables us to extract knowledge from archival data without departing from conventional real-world wisdom. The research employs techniques and theories within the broad fields of mathematics, statistics, and information technology, including signal processing, probability models, machine learning, statistical learning, computer programming, data engineering, pattern recognition and learning, visualization, uncertainty modeling, data warehousing, and high performance computing.

Constructing TIPD big archival data requires a set of skills, such as back-end programming skills (e.g. Assembly, C, C++, Pascal, Delphi, Java, etc.), integration of algorithms, processing big data using distributed data storage and processing systems (e.g. Hadoop Map/Reduce etc.), manipulation of structured data (e.g. SQL, JSON, XML) and unstructured data (e.g. NoSQL, text mining), data manipulation and processing tools (e.g., Python, R, SAS, etc.), web programming skills (e.g. JavaScript, HTML, CSS), systems administration, math (e.g. linear algebra, real analysis, calculus), numerical optimization (e.g. line search algorithms), classical statistics (e.g. general linear models, experimental design, discrete data analysis, etc.), Bayesian statistics and Monte-Carlo simulation, machine learning (e.g. decision trees, neural nets, SVM, classification, etc.), temporal statistics (e.g. time-series analysis), spatial statistics (e.g. geographic covariates, GIS), graphical models (e.g. social networks, Bayes networks), simulation (e.g. agent-based modeling, micro-macro link modeling), visualization (e.g. statistical graphics, mapping, web-based visualization), business, surveys and marketing management.

C. Overcoming Legal and Ethical Issues

Legal and ethical issues are a top priority in this research. All computing tasks are conducted in a closed, supervised data lab of the government funding agency (1) to extract valuable info embedded in confidential micro data of the household registration system, and (2) to enrich extracted info through the processes of cleaning, cleansing, crunching, reorganizing, and reshaping the source data to produce a number of data sets that contain no individual info and thus can be opened to the public to promote open administrative data analytics study. To construct TIPD big archival data, the research constructs automated data processing procedures.

We now turn to open data, using TIPD as an example. Constructing open data is not a goal of the joint research program at its initial stage. However, one crucial question is: why does the joint research program bother to construct TIPD? It all comes from concerns about privacy, confidentiality, legal, and ethical issues. Because only the PI is allowed to access the micro individual data sets and not all research team members are specialized in coping with complex issues of raw data and/or in conducting scientific computing, it becomes urgent to design a way that allows massive raw data sets to be processed and transformed to a set of data in a systematic and automated way.

The transformed new sets of data must fit two criteria: first, they must preserve the main features and most information embedded in raw data; second, they must resolve privacy, confidentiality, and thus legal issues; third, they must comply with academic research ethical requirements. With such issues revolved, the constructed data sets fit the criteria of open data and thus can be utilized directly by the research team members. Because
the constructed open TIPD data sets have been proven effective in promoting the efficiency of the joint research program, the author thus decided to open TIPD to the public, a ray of hope in promoting efficiency, collaboration, mutual trust, and transparency in Taiwan Indigenous Peoples studies. That is why “CopyLeft(L)” is highlighted as a main feature of TIPD.

To create open data, the author decided to adopt “old-school” multi-dimensional tables (MDTs), as illustrated in Figure 1, as a simple but effective alternative to protect privacy while keeping the embedded information of source data sets nearly intact. The foundation of MDTs essentially resembles that of modern distributed storage/processing systems like Google’s, Apache Hadoop, etc. [8] [9].

The main reasons for adopting the MDT method are twofold. First, the learning curve of modern distributed file/storage/computing systems is very demanding for team members. Second, constructing TIPD does not require such a complex system. Fortunately, by reviewing the foundation of the aforementioned contemporary distributed file/storage/computing systems, the author found that their foundation resembles that of classical “old-school” multi-dimensional tables that have been used by, e.g., Statistics Canada and the US Census Bureau for a very long period. As a result, the research adopts conventional multi-dimensional tables as a means for “distributed data storage” and “centralized data integration”.

D. Computational Infrastructure

The infrastructure of computational archival big data is illustrated in Figure 2. Regarding data processing, enrichment, and sharing, the processes of cleaning, cleansing, and reorganizing raw data consist of the following computing work: dealing with distributed data sources, tracking data provenance, error-checking raw data, validating data, coping with missing values and heterogeneity, working with different data formats and structures, ensuring data integrity and data security, enabling data discovery, integrating raw data, and developing algorithms that exploit parallel and distributed architectures to process and enrich the content of raw data.

The aforementioned processes of raw data manipulation are mainly achieved by a set of computer programs that are coded with object Pascal (or Delphi) programming language. Each program is coded and compiled to be run in command line mode. We are thus allowed to make use of redirection and pipe line functionality to integrate different programs and to process raw data in a systematic and automated way through batch processing.

III. METHODOLOGY

A. Digital Infrastructure

Hardware infrastructure does matter and plays a very crucial role in constructing TIPD, particularly in scientific computing for probabilistic record linkage [10]. Because the construction of TIPD is not simple sequential data processing, the research adopts the following principles to manipulate settings of the digital infrastructure (including data streaming setting, disk-based setting, distributed computing setting, and multi-threaded setting) to enhance computing efficiency. First, CPUs, DRAM, and supporting BIOS on motherboard must allow manipulation for acceleration (so-called “overclocking”). Such a setting ensures that in-memory computing functionality is allowed. Second, always utilize as many CPU cores as possible in order to reduce computing time. Third, load as many data sets, including intermediate temporary data during processing, into memory. Fourth, make use of a simple RAID0 setting of high-end SSDs as
a cache for a temporary data repository, and accelerate the motherboard’s chipset via tuning BIOS settings to speed up data transfer between CPUs and storage devices such as RAID0 SSDs.

B. Computational Methods

The research adopts a high performance workstation that has two high-end Xeon 2680 v2 CPUs and 256 GB DDR3-16000 ECC DRAM and a BIOS which allows us to adjust I/O bus and internal memory information transferring speed. In order to control and take full advantage of digital hardware settings that enable us to save a substantial amount of computing time, the author has developed computing codes in object Pascal programming language and has had the codes compiled by an Embarcadero RAD Studio v. 10.1 compiler.

IV. TYPES OF CONSTRUCTED ARCHIVAL BIG DATA

A. Linked Archival Data: genealogy and population dynamics

Record linkage, including deterministic and probabilistic linkage, is a conventional way to extend the value of data [11]. This research utilizes record linkage methods, including both deterministic and probabilistic ones, to create two broad categories of data: one is population dynamics data (see Figure 4) and the other genealogy data. It is worth stressing that in-memory computing serves as a very crucial method to construct TIPD. The computing tasks for creating TIPD involve complex record linkage, including both deterministic and probabilistic matching. This is particularly important, for example, while constructing micro genealogy data by record linkage between a master databank and reference databank through the links of parents’ names and spouse’s name.

To reduce unnecessary searches in the reference databank, the reference databank is sorted by the order of gender, family name, and given name. Furthermore, the sorted reference databank is indexed by a file that records information on the first row of record for each sequence of gender, family name, and given name in the sorted reference databank [10]. Taking the construction of individual genealogy for example, the research matches each individual in the master databank with the reference databank, with respect to the individual father’s name, mother’s name, and spouse name. Since the average number of searches in the sorted reference databank for each record linkage by name is about 50 thousand, the total number of searches involved in matching records of parents and spouse amounts to around 81 billion times.

The record linkage between the master databank and the sorted reference databank is implemented by the following procedures: first, for any given individual record in the master databank, use information stored in

Figure 3. Acceleration of digital hardware: CPUs, DRAM, SPD, and RAID0 SSDs

(a) Overclocking CPUs
(b) Overclocking RAM Disk
(c) Overclocking RAID0 SSDs

Figure 4. Computational process of constructing population dynamics archives
the index file to acquire the first row of information for the name to be matched in the sorted reference databank; second, use information retrieved from the index file to locate the first record in the sorted reference databank; third, search the name to be matched in the sorted reference databank; if the name to be matched is not unique in the sorted reference databank, we choose the record with the maximum likelihood as the person to be matched using information on age and ethnicity; fourth, pick up the matched individual record in the sorted reference databank and insert it at the end of the individual for matching in the master databank. To accelerate the construction of a micro databank of human relationships and kinship, the research takes advantage of in-memory high performance computing (HPC) techniques.

B. Geocoding Individual and Household Data

The research locates indigenous peoples through locating households based on address information. The research at the initial stage used the Taiwan Address Matching System (TAMS) to locate indigenous peoples’ households starting in 2013. However, it was found that the Google geocoding system provides us with more precise and accurate geocoding information, so the research changed the geocoding tool from TAMS to Google’s geocoding system in early 2015. As a result, the geographic coordinate system used to locate households of Taiwan indigenous peoples is WGS84.

The total number of indigenous households amounts to around 210,000. The research has managed to locate 97.8% of indigenous households, with the geocodes of the remaining indigenous households being projected by using information of adjacent indigenous households that have been successfully located. Based on fully located indigenous household geocodes, the research then went further to locate Taiwan indigenous peoples from different perspectives of population geography. The research finished by locating the following indigenous peoples: (1) indigenous households (about 210,000 in number); (2) each indigenous individual (about 560,000 persons in total).

Based on the WGS84 geocodes of a total of 560 thousand indigenous individuals, the research then went further to locate indigenous population centers by different geographic units. The located population centers of indigenous peoples include (1) four regional indigenous population centers; (2) 25 prefecture and city population centers (equivalent to county level); (3) 365 township indigenous population centers; (4) 8,700 village/Chun-Li indigenous population centers. On the basis of the WGS84 geocodes of a total of 210,000 indigenous households, the research managed to locate 486 indigenous tribes and communities by ethnic groups. The results of locating Taiwan indigenous peoples and four population grouping levels (traditional tribe/community, Chun-Li, township, and prefecture/city) are visualized as maps shown in Figure 5.

![Figure 5. Population centers of Taiwan Indigenous Peoples](image)

C. Types and Repository of TIPD Big Data

The aforementioned four population centers are aggregated from the whole data set of individual geocodes, based on the WGS84 projection system. Because existing literature stresses that the spatial distribution of ethnic groups is crucial to explain ethnic in-group organization and ethnic inter-group interaction, the author demonstrates the maps of TIPs’ individual spatial distribution, with respect to each ethnic group.

Major outputs of TIPD which are open to the public amount to 38,000 files and around 85 GB in size as of October 2017. TIPD is bilingually documented, and its contents, context, and volume are growing steadily. TIPD data repository is hosted by Open Science Framework (https://osf.io). For details, please refer to https://osf.io/e4rvz/ (DOI 10.17605/OSF.IO/E4RVZ, ARK c7605/osf.io/e4rvz). Main outputs of TIPD applications include cross-sectional categorical data, longitudinally constructed population dynamics data, life tables, household statistics, micro genealogy data, intra- and inter-ethnic marriage data, ethnic integration data, ethnic patriarchy and matriarchy identity data, etc.

For example, Figure 6 illustrates ethnic marriage practice by generations and by gender that is derived from constructed genealogy archives [10]. Information revealed by Figure 6 suggests (1) that contemporary TIPs are associated with much more exogamy practice than their
parents; this is particularly noteworthy for the ethnic Amis, which is the largest ethnic group among TIPs; (2) that intra-ethnic endogamy is prevalent in the four largest ethnic groups; and (3) that female TIPs practice more exogamy than their male peers. Based on TIPD big archival data, the formation of ethnic identity among TIPs has finally been unveiled [10]. The research identifies 97% of TIPs’ individual ethnic identity formation, with 50.6% of TIPs’ ethnic identity originating from mono-ethnic identity, 19.4% from patrilineal-ethnic identity, and 27.0% from matrilineal-ethnic identity.

TIPs are ideal for studies on issues related to ethnic identity and ethnic marriage as well as social cohesion and wellbeing [10]. By using TIPD, ethnic marriage and ethnic identity formation are associated with distinct patterns between parental and child generations, and between males and females. In addition to the effects of generation and gender, the research also finds that education and spatial organization are crucial in explaining ethnic marriage practice. The general findings include that parental marriage practice is mainly dominated by endogamy; that contemporary TIPs practice exogamy much more than their parents; that intra-ethnic endogamy is prevalent in the large ethnic groups; and that female TIPs practice exogamy more than their male peers.

TIPD consists of three categories of open research data: (1) categorical data, (2) household structure and characteristics data, and (3) population dynamics data. Categorical data include two broad dimensions. The first one is contingency tables which are available in PDF, HTML, RTF, XLS formats, while the other is multi-dimensional data which are offered in CSV, Excel, dBase, Access, Matlab, Gauss, HTML, JMP, SAS, SPSS, Stata, and R formats. Household structure and characteristics data consist of three broad dimensions of information: (1) household head information, (2) household member/composition information, and (3) household geographical information. They are also available in CSV, Excel, dBase, Access, Matlab, Gauss, HTML, JMP, SAS, SPSS, Stata, and R formats.

Another innovative outcome of TIPD are population dynamics data. Population dynamics data consist of three categories: (1) increased population within a given period of time, which can be further dichotomized into two branches of data, population increase due to birth and that due to immigration; (2) decreased population within a given period of time, divisible into two branches of data, population decrease due to death or that due to emigration; (3) intact population within a given period of time. This can be distinguished into two categories of population: those who migrate internally and those staying put. For intact population with internal migration, internal migration processes such as in-, out-, net, and gross migrations are analyzable.

TIPD also offers additional open data for specific research purpose, including (1) indigenous tribal data in traditional territories, (2) indigenous communities in urbanized areas, and (3) aggregated population centers at various geographic units.

D. Practices of Data Analytics

Similar to other countries’ ethnic minority population, contemporary TIPs in Taiwan are associated with lower SES, much shorter life expectancy, and more disadvantaged labor market outcomes such as income gains, employment opportunities, etc. [12]. Although TIPs share of the total Taiwan population is only 2.3%, the importance of research on TIPs lies in the following facts. Based on the author’s previous co-authored studies on the internal migration of TIPs, TIPs are characterized by four features in terms of population distribution and migration: (1) geographically segregated population distribution, (2) very migratory and mostly rural-to-urban migration, (3) periphery of metropolitan areas serving as main destination choice for TIPs rural-to-urban migrants; (4) weak ability of TIPs migrants to engage in onward migration, with most choosing return migration; once repeat migration occurs, TIPs are more likely to choose return migration [13].

TIPD offers rich data to be explored. For example, descriptive statistics are easily available based on TIPD regarding TIPs population size, percentage of population residing in metropolitan areas, gender, age, education, and marital status with respect to each ethnic group of TIPs [10]. The four largest ethnic groups (Amis, Atayal,
Paiwan, and Bunun, see Figures 7.b, 7.c, 7.d, & 7.e, respectively) make up 81% of the total TIPs population, with ethnic Amis constituting 37% of TIPs. Except for ethnic Amis, who reside mostly either in eastern Taiwan or metropolitan areas of northern Taiwan, most ethnic groups of TIPs have fewer members residing in metropolitan areas. Sex ratio (the ratio of male to female population) demonstrates that except for two small ethnic groups (Sediq and Sakizaya, see Figures 7.1 & 7.m), female TIPs outnumber males in terms of total population size. This sex ratio is particularly noteworthy for ethnic Atayal, and is predominantly shaped by two factors: first, male TIPs have a much lower life expectancy than the national average for males; second, the life expectancy of female TIPs does not differ too much from the national average level of female life expectancy.

In terms of educational composition as measured by the percentage of the population with at least some college, ethnic Thao (see Figure 7.o) and Sakizaya have the highest rates (20.5% and 18.4%, respectively), and it is not surprising to find that ethnic Dao (see Figure 7.n), who mainly reside on a small island, have a rate of only 9.3%. The age structure of TIPs is characterized by a relatively low proportion of elderly individuals (aged 65 and over) and a larger young population segment (aged 15 and less), with ethnic Sakizaya as an exception [10].

Individual geocoding results in TIPD help unveil embedded and interwoven social networks of TIPs. For example, Figure 7 (shown in the last page of this paper) demonstrates that the overall pattern of TIPs’ population spatial distribution is characterized by three features: (1) rural TIPs mainly reside in eastern Taiwan and in central mountainous highlands, (2) urban TIPs are seen to concentrate mostly in northern Taipei and Taoyuan metropolitan areas and partly in southern Kaohsiung and central Taichung metropolitan areas, and (3) TIPs barely reside in the rural areas of western Taiwan, which have been mainly populated by ethnic Han and other ethnic groups since the 17th century.
associated with a lower share of population residing in metropolitan areas.

V. CONCLUSION AND DISCUSSION

TIPD reflects the progress and efforts of Taiwan academicians struggling to construct various developments of contemporary TIPs population. The potential applications for research on TIPs based on TIPD include studies on birth process, death process, migration process, residential mobility, life tables, marriage, ageing, education, medical care, labor, family, community, etc. TIPD could be used as background data for survey study, including population analysis, sampling design and sampling planning. Not only does TIPD help unveil contemporary TIPs demographics and various developments, but it also helps overcome research barriers and unleashes creativity for TIPs studies. It will help shed light on contemporary population, human dynamics, and developments of TIPs which have been “invisible” to the world for seven decades.

Potential contributions: TIPD’s potential contributions are threefold. First, theoretically based on data science, not only does TIPD overcome legal and ethical issues, but it also democratizes the use of detailed information hidden in modern micro data sets. Thus it is expected to promote research and unleash creativity in the context of TIPs studies and to enhance the visibility of TIPs. Second, TIPD empirically demonstrates that value-added data enrichment and open data sharing can be accomplished by using less expensive digital infrastructure and an open data repository. Third, in addition to general-purpose research, TIPD enables us to conduct very specific research, such as population dynamics, family life course, life tables, ethnic relationships, etc.

In the past decade, we have seen that the emergence of high-performance computing and supercomputing is reshaping research methods in the social sciences. Computation exploring social complexity based on simplicity like the construction of micro human relationships and kinship in the present research is becoming feasible. As this research demonstrates, computing issues remain challenging, and total costs of computing are still time expensive. The emerging data science, which integrates multi-disciplinary skills and knowledge of manipulating digital infrastructure, expertise in advanced mathematics and statistics, and domain knowledge, is proven to be a crucial new way, or interdisciplinary study, that enables us to overcome some conventional research constraints.

In short, the potential contributions of TIPD are as follows. First, a contribution of moving from “closed” to “open” in the sense that the research on TIPD helps shed light on contemporary Taiwan Indigenous Peoples and human dynamics which have been “invisible” to the world for seven decades. Second, a contribution of moving from “the elite” to “the ordinary” in the sense that the constructed open data sets reduce tech barriers for
Researchers interested in indigenous population studies. Third, a contribution of moving from “local” to “global” in the sense that the English version of TIPD is open to the international academic community to promote further value-added data enrichment through international collaboration. Fourth, a contribution of enabling TIPs research from “macro and static” to “micro and dynamic” data by providing, e.g., micro social network data, genealogy, and population dynamics open data.

ACKNOWLEDGEMENT

The research is mainly supported by a grant from the 2013-2017 bilateral joint research program of Academia Sinica and the Council of Indigenous Affairs on “Research on Contemporary Taiwan Indigenous Peoples”, and partly by the Ministry of Science and Technology (MOST 106-2420-H-001-008-MY2). The author appreciates the excellent research assistance from Hin-Sen Liao and Amy M.L. Chen.

REFERENCES