PROCEEDINGS OF SCIENCE

Data Science as a Foundation towards Open Data and Open Science: The Case of Taiwan Indigenous Peoples Open Research Data (TIPD)

Ji-Ping Lin[1]
RCHSS, Academia Sinica
128, Sec. 2, Academia Rd., Nankang, Taipei, Taiwan
E-mail: jplin@sinica.edu.tw

The research is an outcome of the joint research program by Academia Sinica and the Council of Indigenous Affairs in 2013-2017. The aims of this paper are threefold: (1) to demonstrate the methods of data science in constructing the Taiwan Indigenous Peoples (TIPs) Open Research Data database (TIPD, see http://TIPD.sinica.edu.tw, and https://osf.io/e4rvz/, identifiers: DOI 10.17605/OSF.IO/E4RVZ, ARK c7605/osf.io/e4rvz) based on Taiwan Household Registration (THR) administrative data; (2) to illustrate automated and semi-automated data processing as methods for constructing effective open data; and (3) to demonstrate appropriate utilization of "old-school" data formats such as multi-dimensional tables as an effective means to overcome legal and ethical issues. The research extracts valuable information embedded in micro data of THR and enriches the extracted information through the processes of cleaning, cleansing, crunching, reorganizing, and reshaping the source data. The data enrichment processes produce a number of data sets that contain no individual information but retain most of the source data information. The enriched data sets thus can be open to the public as open data. The open data are systematically constructed mainly in an automated and partly in a semi-automated way through the integration of optimized hardware, compiler & script programming languages, computing software, and system script languages. Major outputs of TIPD amount to 31,000 files in number, totaling around 79 GB in size. TIPD consists of three categories of open research data: (1) categorical data, (2) household structure and characteristics data, and (3) population dynamics data. The potential contributions of TIPD are moves from "closed" to "open" , from "the elite" to "the ordinary", from "local" to "global", and from "macro and static" to "micro and dynamic" research.

Keywords: big data, data science, open data, open science, TIPD, TIPs

International Symposium on Grids and Clouds 2017 -ISGC 2017-
5-10 March 2017
Academia Sinica, Taipei, Taiwan

[1] Speaker

PoS(ISGC2017)004

1 Introduction

Taiwan Indigenous Peoples (TIPs) are a branch of Polynesian-Malaysian (or Austronesian) ethnic groups in the genetic and linguistic contexts. Since the early 17th Century, TIPs played a crucial role during the Great Marine Times of East Asia trades [1] [2]. There was a rich body of ethnographic, official and academic records on TIPs before 1940. However, the period of 1940-2000 was the data "Dark Ages" for TIPs due to the 1941-45 Pacific War, and 1946-1990 political authoritarian rule due to fears of communism and communist infiltration. Persistent lack of TIPs data led TIPs to become isolated, marginalized and thus underdeveloped.

Taiwan resumed the TIPs population census in 2000 and began recording TIPs' individual records in the Taiwan household registration system (THRS) in 2003. After a decade of efforts, TIPs records in THRS have become complete and much more consistent. This research program is conducted on the basis of a four-year joint research agreement between Academia Sinica and the Council of Indigenous Affairs starting in 2013. One important aim of the research is to construct big anonymous TIPs open research data (or Taiwan Indigenous Peoples open research Data, TIPD thereafter in the paper) based on contemporary census and household registration data sets. TIPD utilizes state-of-the-art data science, record linkage, geocoding, and high-performance in-memory computing technology to construct various dimensions of TIPs' demographics & developments.

This article uses TIPD to manifest how data science serves as a foundation for open data and open science. The research aims to demonstrate how data science, i.e., integration of hacking skills, advanced math/statistics knowledge and skills, and domain knowledge expertise, is applied to construct Taiwan Indigenous Peoples open research Data based on Taiwan Household Registration administrative micro data. As shown in Fig. 1, detailed information about TIPD can be found at this joint research program website at http://TIPD.sinica.edu.tw.

There are three elements we need to highlight to distinguish their relationship at first. They are, in short, data science, open data, and open science. Open data refer to data that can be available to everyone to use and republish without restrictions of any form, including copyright, patents, and/or other mechanism controls. The spirit and goals of open data resemble those of other "open" movements like open source, open hardware, etc.[2] A common misunderstanding about open data among the ordinary people is that open data do not place restrictions to protect privacy and to preserve confidentiality, leading to legal and ethical issues in the end. Such misunderstanding resembles the misunderstanding and misconception during the early phase of

[2] For reference, see https://en.wikipedia.org/wiki/Open_data

the open source movement in the sense that open source will violate copyright and thus produce legal issues. As a matter of fact, emphases on protecting privacy and preserving confidentiality are the first requirement of the open data movement.

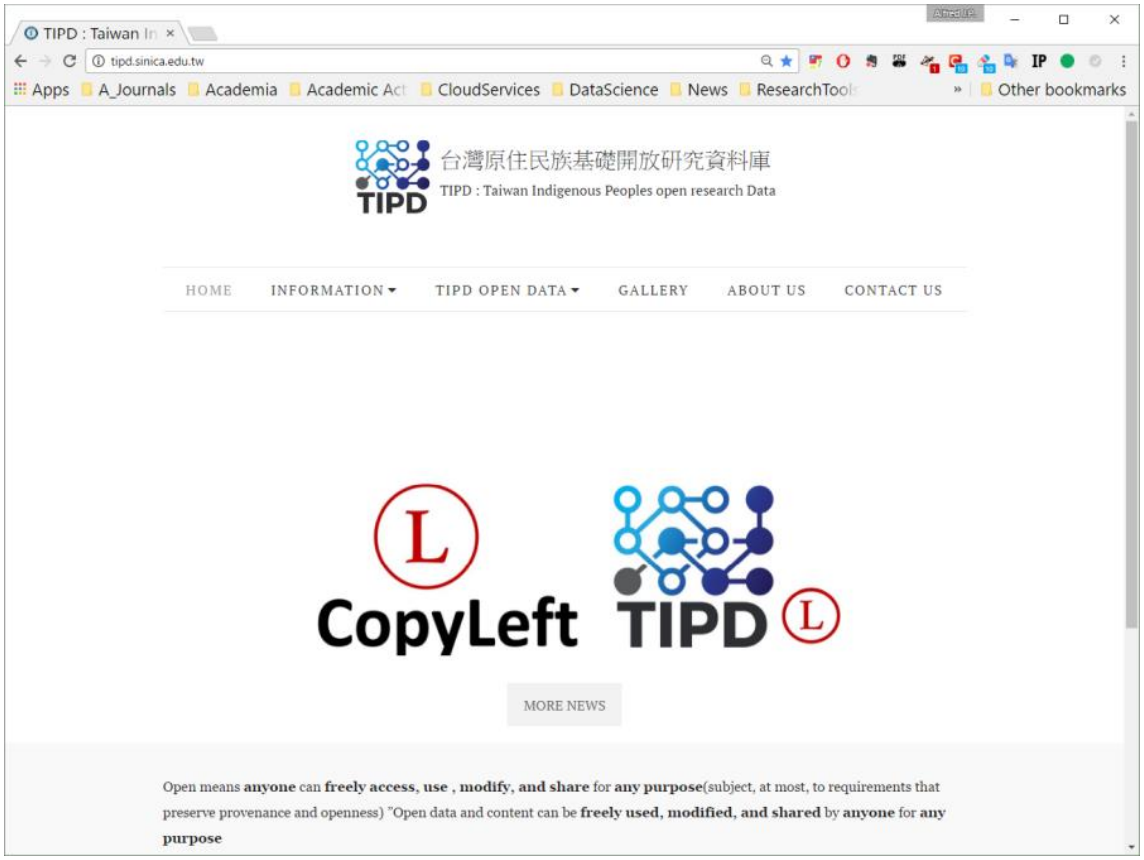

Fig. 1, Taiwan Indigenous Peoples Open Research Data (TIPD) at https://TIPD.sinica.edu.tw

The calls for "Open Science" emerge due to the flaws of "Closed Science", e.g., issues of scientific reproducibility, and access to data and knowledge systems. One important feature in open science is "replicability" and "reproducibility" given the same methods and data. However, there have been rising concerns in academic communities about scientific replicability and thus research credibility issues in the past few years. One result is the call for open science. By definition, open science is the movement to make scientific research methods, data, and results accessible to all communities. Open science consists of six principles: open data, open source, open methods, open peer review, open access, and open educational research.[3]

We now turn to open data, using TIPD as an example. TIPD is the largest open academic research data set in Taiwan now. The research program funding period was from 2013 to 2016, which has been extended for one year in 2017. In only a couple of years, TIPD has contributed to helping overcoming data issues and promoting contemporary Taiwan indigenous peoples

[3] For reference, see https://en.wikipedia.org/wiki/Open_science

research. Although it is a by-product of the joint research program on Rural-to-urban Taiwan Indigenous Peoples, TIPD is the most important and influential output of this joint research program. Constructing open data was not a goal of the joint research program at its initial stage, so why did the joint research program bother to construct TIPD? It all comes down to concerns on 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 became 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 the raw data; second, they must get rid of privacy, confidentiality, and thus legal issues; third, they must comply with academic research ethical requirements. Given the aforementioned situations being fulfilled, i.e., legal and ethical issues being solved, 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 data sets proved to be effective in promoting the efficiency of the joint research program, the PI 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.

This research reports the progress and efforts of Taiwan academicians struggling to construct contemporary TIPs Data (namely, TIPD) by integrating the micro data of the 2000 Taiwan population census and household registration system, using state-of-the-art data science technology like geocoding, record linkage to distinguish natural and social increases, using high-performance computing (HPC) to construct micro data of human relationships and kinship to study patterns of inter- and intra-ethnic marriage and levels of integration, etc. This research demonstrates the foundation of data science in processing and enriching administrative data. Central to the construction of TIPD is data science. The role of data science in constructing TIPD is discussed in Section 2. Methodology is highlighted in Session 3. In Section 4, the paper makes a brief introduction on TIPD research results, including data contents and potential applications. Section 5 offers concluding remarks.

2 The What, Why, and Role of Data Science

Just as the term *Big Data* started becoming popular a few years ago, the term *Data Science* is gradually gaining more attention in academic communities and industries [3] [4]. Nevertheless, it is worth stressing that data science is by no means a

new field of science. Rather, it is multidisciplinary in essence. There are various definitions of data science [5] [6]. The simplest but most informative is the one that defines data science as a scientific framework that consists of three necessary components: (1) hacking skills, (2) advanced mathematics and statistical knowledge and skills, and (3) domain knowledge expertise, as shown in Fig. 2 [5]. Data science is based on real world domain knowledge expertise and refers to the extraction of knowledge from data, with the main goals of enhancing human knowledge from data and producing data products. It 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.

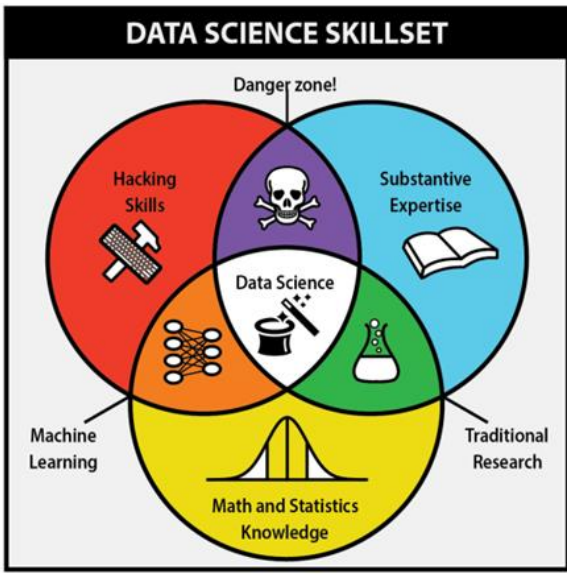

Fig. 2. A definition of Data Science (source: Source*: [5])*

It is worth stressing that hacking skills have nothing to do with illegal hackers' activities. Hacking skills refer to the skills and ability of manipulating and integrating digital infrastructure that includes hardware settings, operating systems, programming languages, and software. Data science 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 and text mining), data manipulation and

processing tools (e.g., Python, R, SAS, etc.), web programming skills (e.g. JavaScript, HTML, or CSS), systems administration, math (e.g. linear algebra, real analysis, and 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 and GIS), graphical models (e.g. social networks and Bayes networks), simulation (e.g. agent-based modelling or micro-macro link modeling), visualization (e.g. statistical graphics, mapping, or web-based visualization), business, surveys and marketing management [7] [8].

It has been recognized that issues corresponding to administrative data are quite different from those associated with survey data [9]. It is also accepted by many academicians that researchers have benefitted from richer and more reliable sources of administrative data that provide us with deeper insights and much broader vision than survey data about policy and socioeconomic behavior. Conventional administrative data gathered for administrative purposes by government agencies essentially resemble contemporary big/massive micro data gathered by industry for commercial purposes, although the data collection media tend to differ from each other. Because data science is playing a crucial role in the contemporary data revolution in industry, it is worth noting that little attention has been paid in the sense that methods of data science might have great potential to tackle administrative data in terms of data collection, parsing, cleaning, cleansing, validation, privacy-protection, quality assessment, reorganization, processing, exploration, analysis and enrichment issues [9] [10].

3 Methods

The raw data sets of TIPD mainly come from Taiwan Household Registration Data (THRD) and the 2000 Taiwan population census. Main variables in the raw data include: household ID, PIN, name, spouse's name, father's name, mother's name, address, gender, birth date, marital status, education, prefecture/city code, township code, village code, birth place code, relationship with household head, and ethnic group code. Each raw data set was collected on a bimonthly basis during 2013 and June of 2015 and on a monthly basis after July 2015. By the end of 2016, the total number of raw data sets amounted to 34.

The processes of cleaning, cleansing, and reorganizing the 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 processes 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. In such a situation, 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 the procedures of batch processing.

In terms of data integration, repository, and sharing, the research adopts the following simple but effective methods. First, open data integration is achieved by batch processing on a local computer. The research makes use of Dropbox and Google Drive as data repository platforms for the integrated open data sets. The contents of TIPD on Dropbox and Google Drive are always identical to each other through a standardized process of data synchronization with a local data repository. To share research information and TIPD open data, the research does not set up a data-sharing server; rather, after a very careful evaluation of open data repository options, the research adopts a Nature-recommended Open Science Framework (OSF, see https://osf.io/) platform. Because the research grants OSF access to the TIPD repository on Dropbox and Google Drive, any individual is thus allowed to browse open research information of TIPD and to download any data sets of TIPD via OSF.

Hardware infrastructure does matter and plays a very crucial role in constructing TIPD, particularly in scientific computing for probabilistic record matching [11]. 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, the CPUs, DRAM, and supporting BIOS on the motherboard must be manipulated for acceleration (the so-called “overclock”). 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, as possible 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 speeds between CPUs and storage devices such as RAID0 SSDs.

Now let's turn to methods used to overcome legal and ethical issues. First of all, the research gives up the conventional in-house data lab as the way to protect privacy and confidentiality. Not only is the in-house data lab mode very inefficient (e.g. you are not allowed to customize computing environment settings or use your own computing facilities), but it is also very expensive to use (e.g. open only during office hours and needing daily commuting to an in-house data lab). Although contemporary real-world distributed storage file systems and computing environments (e.g. Google's distributed file/storage/computing system, Yahoo Hadoop, and Apache Spark) are ideal to overcome legal and ethical issues, it is not practical to adopt such settings to construct TIPD. The main reasons are twofold. First, the learning curves of modern distributed file/storage/computing systems are 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 PI found that their foundation resembles that of classical "old-school" multi-dimensional tables that have been used by, e.g., the Statistics Canada and US Census Bureau, for a very long period. As a result, the research adopted conventional multi-dimensional tables as a means for "distributed data storage" and "centralized data integration".

Record matching is a conventional way to extend the value of data [12]. 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 matching, including both exact and probabilistic matching. This is particularly important, for example, while constructing micro genealogy data by record matching 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 [11].

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 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 the matched individual record in the sorted reference databank and insert it at the end of the individual record for matching in the

master databank.

Based on the above mentioned record searching and matching procedures, 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's name. Since the average number of searches in the sorted reference databank for each record matching by name is about 50 thousand, the total number of searches involved in matching records of the parents and spouse amounts to around 81 billion times. 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. In-memory high-performance computing comprises three kernel skills of manipulating digital infrastructure: (1) overclocking CPUs, (2) overclocking internal memory speed, and (3) accelerating I/O bus bandwidth that links CPUs and internal memory.

The research adopts a high performance workstation that has two high-end Xeon 2680 v2 CPUs, 256 GB DDR3-16000 ECC DRAM, and a BIOS which allows us to adjust I/O bus and internal memory information transfer speed. In order to control and take full advantage of digital hardware settings that enable us to save substantial amounts of computing time, the author developed computing codes in the object Pascal programming language and had the codes compiled by an Embarcadero RAD Studio XE6 compiler. The programming codes are designed for in-memory computing purposes in the sense that all computing tasks of constructing human relationships and kinship databank are implemented in computer's internal memory, with CPUs and internal memory being overclocked and I/O bus between CPUs and memory being accelerated.

4 Results

Major outputs of TIPD which are open to the public amount to 31,000 files in number and around 79 GB in size. TIPD is bilingually documented, and its content, context, and volume are growing steadily. The TIPD data repository is hosted by the Open Science Framework (https://osf.io), as shown in Fig. 3. 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- & inter-ethnic marriage data, ethnic integration data, ethnic patriarchy and matriarchy identity data, etc.

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 is contingency tables which are available in PDF, HTML, RTF, and 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 Access 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 Access formats.

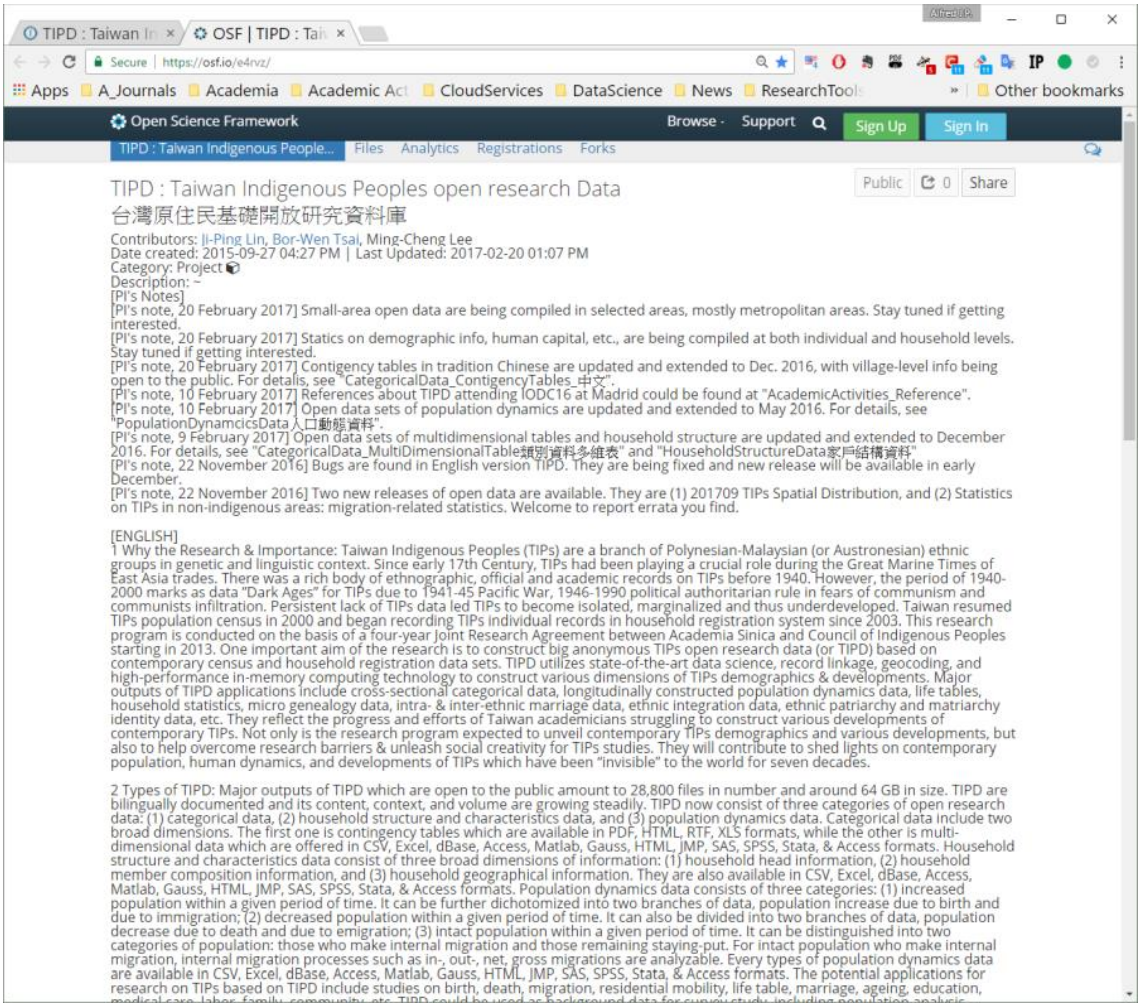

Fig. 3. TIPD data repository on Open Science Framework (OSF)

Population dynamics data consists of three categories: (1) increased population within a given period of time, which can be further dichotomized into two branches of data, population increases due to birth and due to immigration; (2) decreased population within a given period of time, which can also be divided into two branches of data, population decreases due to death and due to emigration; (3) intact population within a given period of time, which can be distinguished into two categories of population: those who make internal migration and those

stay put. For an intact population with internal migration, internal migration processes such as immigration, emigration, net migration and gross migration are analyzable. Every type of population dynamics data is available in CSV, Excel, dBase, Access, Matlab, Gauss, HTML, JMP, SAS, SPSS, Stata, and Access formats.

5 Concluding Remarks

The construction of TIPD reflects the progress and efforts of Taiwan academicians struggling to understand various developments of contemporary TIPs. The research program is expected not only to unveil contemporary TIPs demographics and various developments, but also to help overcome research barriers and unleash social creativity for TIPs studies. These will shed light on contemporary populations, human dynamics, and developments of TIPs which have been "invisible" to the world for seven decades. The potential applications of TIPD to TIPs studies include studies on birth, death, migration, residential mobility, life tables, marriage, ageing, education, medical care, labor, family, community, etc. TIPD could be used as background data for survey studies, including population analysis, sampling design and sampling planning. As R has becoming a common language among data scientists, the author will use the R data format for the project when releasing the next waves of TIPD data.

Potential contributions: TIPD 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 the 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.

With data science as a foundation to construct TIPD, the research achieves the following goals that may contribute to open science: (1) designing a data enrichment process that allows us to construct open data in an automated and consistent way by using high-performance computing digital infrastructure; (2) reorganizing raw data as open data to overcome legal and ethical issues based on the foundation of distributed storage methods; (3) allowing us to enrich data through the process of data integration methods, making longitudinally linked data (including both exact and probabilistic record matching) less expensive and more efficient; (4) enabling us to go further to extract very detailed information from raw data, such as the construction of micro genealogies, identity, and ethnic marriage patterns.

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 contributes to an understanding of contemporary Taiwan Indigenous Peoples and human dynamics which has been lacking for seven decades. Second, a contribution of moving from "the elite" to "the ordinary" in the sense that the constructed open data sets reduce technological barriers for researchers interested in indigenous population studies. Third, a contribution of moving from "local" to "global" in the sense that English versions of TIPD are 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, genealogical, and population dynamics open data.

References

[1] Blust, A.R. (1985). The Austronesian homeland: A linguistic perspective. *Asian Perspectives,* 26(1), 45–67.

[2] Bellwood, P. (1991). The Austronesian dispersal and the origins of languages. *Scientific American,* 265(1), 88–93.

[3] Schmidt, E. and Cohen, J. (2013). *The New Digital Age: Reshaping the Future of People*, Nations and Business. 1st Edition. Knopf Inc.

[4] National Research Council. (2013). *Frontiers in Massive Data Analysis*. Washington, D.C.: National Academy of Sciences.

[5] O'Neil, C. and Schutt, R. (2014). *Doing Data Science*. CA: O'Reilly Media, Inc.

[6] Avrim Blum, John Hopcroft, and Ravindran Kannan. (2015). Foundation of Data Science, available at https://www.cs.cornell.edu/jeh/book2016June9.pdf

[7] Cioffi-Revilla, C. (2010). Computational social science. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2, 259-271.

[8] Alvarez, R. M. (2016). *Computational Social Science.* Cambridge University Press.

[9] Herzog, T. N., Fritz J.S, and Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques*. Springer.

[10] Lazer, D., et al. (2009) Life in the network: the coming age of computational social science. *Science*, 323, 721-723.

[11] Lin, J.P. (2017). Human Relationship and Kinship Analytics from Big Data Based on Data Science: A Research on Ethnic Marriage and Identity Using Taiwan Indigenous Peoples as Example, in Cathleen M. Stuetzer et al. (ed.) *Computational Social Science in the Age of Big Data. Concepts, Methodologies, Tools, and Applications*. Herbert von Halem Verlag (Cologne, Germany), in Neue Schriften zur OnlineForschung of the German Society for Online Research (DGOF).

[12] Wood, D. (Ed.). 2011. *Linking Government Data*. Springer.