Beijing City Lab

Zhou Y, Long Y, 2015, SinoGrids: A practice for open urban data in China. Beijing City Lab. Working paper #72

SinoGrids: A Practice for Open Urban Data in China

Yulun Zhou [1], Ying Long [2*]

[1] Department of Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong, China

[2] Beijing Institute of City Planning, Beijing, China

* Corresponding author, longying1980@gmail.com, +86-1366-1386-623

Abstract In the past decade, rapid urbanization and industrialization has taken place in China, and is still making a big difference in cities. Mobile Internet has been widely spread, benefiting the dramatic growth of Web 2.0 applications. Nearly everyone is social media user. And cities are evolving into “Wired Cities” with digital devices installed everywhere, intended to monitor, manage and regulate urban flows. As a result, data explosion has taken place in cities. With advanced techniques on data storage and high-performance computing, big/open urban data has opened up important development opportunities for urban studies, planning practice and commercial consultancy. Urban researchers and planners are eager to make use of these abundant, sophisticated, dynamic, and more-spatially-temporally-detailed data to deepen our understanding of urban form and functions. However, lack of proper data for urban analytics is an often-met dilemma. Open urban data in China is still at an early stage. Very few urban data is freely accessible, due to constraints on data distribution and data holders’ concerns of losing their own advantages. Besides, cities are complex systems, efficient and effective interoperation of multi-source urban datasets is a must to draw reliable conclusions on urban behaviors, but dealing with the heterogeneity between datasets is also another critical challenge, especially for urban planners or government officers, who care about and understand cities, need the support from data analysis, but has little data processing experience. In such context, we initiated SinoGrids (Plan Xu Xiake), a crowdsourcing platform for encouraging the standardization (downscaling), sharing and interoperation of micro-scale urban data in China. On one hand, the downscaling preserves the advantage of original data holder and can thus, provide more sharable urban datasets. On the other hand, the downscaling works as a standardization process, making the datasets more manageable, and also, providing a way for effective and efficient interoperation of multi-source urban data. Technical guidelines and tools of the standardization process are provided and freely accessible to make SinoGrids more easily accepted and applicable. Last but not least, a human participated test was proposed for user performance evaluation of SinoGrids.

Keywords: Open Data, Crowdsourcing, Urban Analytics, Citizen Science, China,

1 Introduction

The world’s urban population grew dramatically from 34% (1960) to 54% (2014), and is estimated to grow further by 1.84% from 2015 to 2020 (World Health Organization 2015). In terms of land use, proposed by the Global Rural-Urban Mapping Project (CIESIN 2010), 3.5 million square kilometers of land on earth has been used for urban development. As the vast urbanization and industrialization has taken place worldwide, deepening the understanding of urban form and functions has always been a focus of urban researchers, planners and commercial consultants. Recently, public concern in China on various urban challenges, e.g. air pollution, has been overwhelming (Enserink et al. 2007), strongly encouraging related urban studies. At the meantime, urban big/open data has opened up important development opportunities for urban-related researches by empowering more spatially- and temporally-detailed, sophisticated, large-scale and dynamic analysis on urban issues. Typical urban data application has been proposed on various topics, including house prices (Huang, 2010), mobile phone use (Tranos et al. 2015), accessibility to healthcare (Aoun 2015), etc. In the area of urban and regional planning, Han et al. (2015) discovered functional zones using bus smart card data and points of interest in Beijing, Long analyzed jobs-housing relationships (Long and Thill 2015) and profiled underprivileged residents (Long et al. 2014) based on open/big urban data and Li (2015) proposed an analytical framework for urban planning based on crowd-sourced data.

In the past decade, the amount of urban data that is involving urban residents' daily life is booming, with the installation of digital instruments for monitoring, managing and regulating urban flows everywhere in cities, the rapid spreading of mobile internet, the popularity of social media and Web 2.0 applications, as well as the development of data storage and distribution techniques. (Kitchen 2014) However, as a matter of fact, compared with the large amount of generated urban data, the amount of accessible urban data is limited. Besides accessibility, Gurstein (2011) proposed the confusion between enhancing citizens' "access" to data and enhancing citizens' "use" to data, emphasizing the gap between access and effective use. The Urban system is complex and always involves interactions between disciplines, which makes abundant, high-quality, accessible and usable data be of extra significance in the area of urban and regional studies. For example, for urban air pollution, interactions between urban morphology, climatology, urban land use, etc. are all involved, thus, a wide range of urban data from various data sources are needed to draw reliable conclusion. In such cases, open urban data should be not only accessible, but also usable, reusable and redistributable (Wilbanks 2014) to further empower the public and enterprises for data applications.

The open data movement has been initiated with an overall intention to make local, regional and national data, particularly publicly acquired data, available in a form that allows for direct manipulation, e.g. cross tabulation, visualization, etc. (Gurstein 2011) Typical successful practices include volunteered datasets, e.g. Open Street Map (Haklay et al. 2008), which provide information of urban form and function as the providers are determining urban morphology with their own activity (Crooks et al. 2014). But currently, open data still faces various critical issues, for example, data dispersion, heterogeneity, and provenance. (Gurstein 2011; Overpeck et al. 2011; Reichman et al. 2011)

China has been experiencing an unprecedented vast and rapid process of urbanization and industrialization (Bai et al. 2014). Cities are changing fast. As a result, urban studies in China has even higher requirements on urban data in terms of spatial-temporal scale, resolution, etc. In this paper, focusing on open data challenge for regional and urban studies in China, the purpose will be in two folds. A brief review of open urban data in the world and China, together with related initial practices, comes first. Then, a new attempt of a crowdsourcing platform for providing more sharable and interoperable basic urban data to empower urban and regional studies in China, the SinoGrids (Plan Xu Xiake, in Chinese, SinoGrids: http://www.beijingcitylab.com/projects-1/14-sinogrids/), is proposed. The main approach is by relieving the benefit conflict between original data holder and data user, and bridging the gap between "access" and "effective use". The Chinese name for SinoGrids is Xu Xiake Plan, in which, Xu Xiake is a famous Chinese geographer and travel writer of the Ming Dynasty (1368-1644). He spent 30 years in travelling all around China and documented his travel. By naming SinoGrids after Xu Xiake, we hope urban data all around China could be open and shared on the crowdsourcing platform.

2 Open urban data in the world and China

The definition of urban data we are using, as the context of discussion in this paper, is quite broad. Urban data is referring to all datasets that would characterize some aspects of urban form and functions through interpretation. (Crooks et al. 2014) And the definition of open urban data, subsequently, is therefore referring to urban data that is openly accessible to the public, including researchers, planners, commercial consultants, local residents, etc.

The conventional urban datasets are mainly based on government efforts, e.g. national censuses and cadastral maps. Also, there are administration records, like approvals of construction events from the planning department, economic development reports from the statistics department, etc. As these conventional data are mainly based on regional statistics generated through sampling, urban studies based on these data are suffering from limitation on spatial-temporal scale and resolution, as well as sophistication. Besides, as conventional government data comes in different forms, e.g. paper-based reports, which requires huge digitization workload for detailed, large-scale and sophisticated urban studies. Recent open/big data efforts have partially changed the conventional situation.

Nowadays, there are various new sources for open/big urban data. The first are official data portals, enabled by recent open government initiatives that opens previously non-accessible data sources to the general public. The second are community-generated big data initiatives, collecting data from mobile phone activities, vehicle trajectories, public transit smart card data, business catalogs, and other smart city programs (Batty 2012). There are overlaps between these two sources.

2.1 Government-generated urban data

Mushrooming online data portals run by governments are the best evidence for the official awareness and willingness of promoting social services and transparency, and empowering urban studies by opening data. For example, under Local Law 11 approved by New York City Council in 2012, requiring all agencies to open their data, the *NYC OpenData* (https://nycopendata.socrata.com/), funded by New York City, provides nearly 1300 datasets as of July 2013, and has a further plan of 345 more datasets to be released before 2018. The included datasets concern various aspects of urban activities, e.g. public safety, city government, education, healthcare and so on. All datasets are in machine-readable formats, paired with corresponding metadata, and some of the data are large in volume. According to U.S. City Open Date Census (Open Knowledge 2014), New York City ranked first in data opening effort in 2014, followed by San Francisco (https://data.sfgov.org/), Los Angeles (https://data.lacity.org/) and Boston (https://data.cityofboston.gov/). At federal level, U.S. Genreal Services Administration also promoted an open data platform, *Data.gov*, providing over 150,000 online datasets gathered from hundreds of organizations including the Federal agencies. In Europe, European Commission has been leading the project, *INSPIRE* (http://inspire.ec.europa.eu/), which is intended to build up a European spatial data infrustructure, and has made spatial data from over 700 data communities open. Detailed technical guidelines of data specifications on spatial elements, e.g. addresses, coordinate reference systems, etc., together with specifications on different categories of data, have been proposed for the mass amount of datasets to be manageable, usable and interoperable.

As China is a developing country, like many others, open data in China faces pressure and limitations from tight regulations, out of national security concerns. Open data in China is still at an early stage. Despite the current situation, the Chinese government has devoted efforts for more data access and more relaxed data control. A national data portal (http://data.stats.gov.cn/) has been initiated by National Bureau

Statistics of China, providing digitized census data, statistical reports on monthly, seasonal and annual scales, as well as some data visualization products. Beijing (http://www.bjdata.gov.cn/) and Shanghai (http://www.datashanghai.gov.cn/), the two metropolises in China, are the first to have established their open urban data platforms, respectively opening over 400 and 209 urban datasets provided from various government departments. Wuhan is also expecting the opening of its “one-cloud, one-map, one-standard, one-model, one-stop” open data platform with 520 datasets, claiming that all government data would be accessible from this platform in the future. Besides, there are more cities coming, e.g. Qingdao, Guiyang, Guangzhou, etc. The opening of local government data has become a trend in China. (Guo 2014) Summing up from the introductions of the government data platforms, the main motivations and intentions of the opening of government data are: 1) Urban big/open data has been regarded as an effort that corresponds to the human-oriented “New Type Urbanization” in China; 2) Urban big/open data is considered as a signature for the improvement of the accessibility and transparency of government, fulfilling the people’s right to know. Open data movement also helps build up a clean, efficient and open-minded image of the government. 3) The governments are aware that, as the largest holder of public data resource, the best way to make the most out of these data, is not by monopolizing, but by utilizing and sharing.

However, for current government open portals in China, it is found that most datasets are still in the tabular format, which needs further preprocessing for data application. Most data are still based on conventional regional statistics, limiting the spatial and temporal scale, sophistication, and details for urban studies. Besides, the limited number of open datasets is still not enough to fulfill needs from urban research and data applications. Government-dominated urban data sharing is still at an early stage in China.

We also note that government-funded research institutes are taking a leading role in opening data in China. The *Geospatial Data Cloud* (http://www.gscloud.cn/), established by the Chinese Academy of Science, is an open data platform for spatial data, e.g. remote sensing images and retrieval products. Also, some preprocessing services, like atmospheric correction, image gap-filling, etc., are also provided automatically online, making the data more ready for application. Also started by the Chinese Academy of Science, the *Data Center for Resource and Environment Science* (http://www.resdc.cn/Default.aspx) is intended to empower sustainable resource and environment studies in China. The datasets are freely accessible for researchers, while the datasets are mostly about physical geography, concerning vegetation, land, terrain, etc.

2.2 Community-generated urban data

Government-generated data fails to capture the characteristics of urban form and function based on the public’s perception, instead of the administration angle (Crooks 2014). In recent years, community-generated urban data, including social media, volunteered datasets, etc., have risen and opened up new research opportunities in urban studies and planning (Batty 2013). One typical example is the *Open Street Map* (http://www.openstreetmap.org/), which is built by a community of mappers that contribute and maintain data about road networks, point of interests, land use types. *Open Street Map*, as a typical open urban data platform, has the following characters, which is shared among community-generated open data initiatives: 1) Community-driven. The open data portal is maintained by a community of mappers based on local knowledge, which also reflects a public perspective of urban space and activity. 2) Explicit. The data from *Open Street Map* are in GIS-based format and has a close correlation with ArcMap, the most widely used platform for geographic information processing, which makes the data highly usable and interoperable. On the contrary, many other open data, like a paper-based urban land use map, cannot be directly used for analysis and needs further digitization and pre-processing, for which, these data are classified as implicit. 3) Usable, reusable and redistributable. Data from *Open Street Map* is freely usable as long as the user credits Open Street Map as its contributor. The data and result generated could be distributed under certain copyright and license specified. In terms of coverage, *Open Street Map* has also covered many places in China, however, due to various limitations, Chinese cities are measured with lower precision and granularity. Despite that, it has still become one of the finest way of retrieving basic urban data in China.

Locally, leading map service providers, like *Baidu Map* (http://developer.baidu.com/map/) and *Amap* (http://lbs.amap.com/), have provided open data service through Application Programming Interface (API), empowering planners and commercial companies with cloud map service. Big data initiates in China have also been trying to face these challenges. *Datatang* (http://www.datatang.com/), was established as the first data trading platform in China, intended to solve the conflict between the benefits of original data holder and data user in a business way. There were over 44000 datasets available online as of May, 2015, concerning a wide range of areas, e.g. semantic analysis, transportation, and healthcare, etc. But most of them are for sale, only a small proportion is freely accessible and usable. Last but not least, *Beijing City Lab* (http://www.beijingcitylab.com/), a virtual research community focused on urban topics in China, has opened 27 datasets characterizing urban China, which are all freely accessible and in proper format for effective reuse. All the datasets are from Open Data online, and donations from researchers both in and out of the community.

Social media is a new comer of community-generated urban data but has already become a significant part. Currently, the most popular micro-blogging service in the world is Twitter with over 284 million users as of December 2014. Twitter messages are mostly in English, as it is not freely accessible in mainland China. In China, as alternatives, similar micro-blogging services are available, such as Sina Weibo, Tencent Weibo, etc. By September 2014, the number of Monthly Active User (MAU) of Sina Weibo has reached 167 million, with an annual growth of 36%. Promoting the spirit of Web 2.0, which is to encourage user-generated content (UGC), micro-blogging has been an indispensable part of urban life and a major way of expressing personal feelings and opinions in China. Various studies have been proposed for analyzing social media data for urban activities based on Twitter. Tumasjan et al. (2010) tracked public opinion by monitoring political sentiment expressed via social media and predicted election results. The spread pattern of news has been looked into (Lerman et al. 2010), and social media data have even been used to predict earthquakes (Sakaki et al. 2010) and stock market performance (Si et al. 2013). But fewer has been done with China's social media data due to the challenging gaps on semantic analysis between English and Chinese/Cantonese.

2.3 Challenges in open urban data in China

Open data still faces many challenges. Challenges for open data has been reviewed in the field of ecology (Reichman 2011) and climate (Overpeck 2011), among which, challenges of data dispersion, heterogeneity and provenance also applies to urban open data (Liu 2015).

Urban data are generated from various data sources, e.g. government-generated data, volunteered datasets, user-generated content, etc. Data from different sources may have overlaps with each other, and each characterizes a certain aspect of urban form and function. Variations in spatial and temporal extent and scale, data formats, etc., make data interoperation difficult. Urban studies always intersect with adjacent disciplines, which would require high consistency in all aspects of data for interoperation. Also, for implicit data, extra efforts of digitization and pre-processing are needed.

One more important challenge that is unique to open data in China was mentioned by Liu (2015), which is the more frequent and unprecedented changes in urban morphology and behaviors, due to rapid urbanization and industrialization process in China. Characterizing the more frequently changing patterns raises higher requirements on spatial-temporal scale and resolution of urban data, in order to support more detailed and sophisticated urban studies.

In China, as government-dominated open data portals are at an early stage, not being comparable with open data infrustructures like *Data.gov* and *ISPIRE*, urban data are mainly open through a community-driven way. However, for community-driven open data, there is a problem that data holders often have hesitations on data sharing. According to our survey, 79.75% of the surveyed claim that they '*often*' meet the situation of lacking proper data. But only 22.14% of all individuals participated in the survey are willing to share data with any other data users in need. 65.20% are willing to share data with coorperators and 46.20% will share data with colleagues or students. 8.23% of the surveyed would rather not share data with others. The main concerns of data holders lie in losing research advantage (58.62%), lacking of sharing credit (57.76%) and limitations on data distributing by the institution or data provider (54.31%). We can see that, hesitations of data holders has been a serious bottleneck for improving the data opening situation in China.

How to balance the benefits of the original data holder and that of data user remains a fundamental challenge for all open data projects, which is directly effecting the motivation of data holders for community-driven data opening.

In such context, we initiated SinoGrids (Plan Xu Xiake), a crowdsourcing platform for encouraging the standardization (downscaling), sharing and interoperation of micro-scale urban data in China. The downscaling is by projecting the original datasets onto a uniform grid.

3 SinoGrids: A crowdsourcing sharing platform for micro-scale basic urban data in China

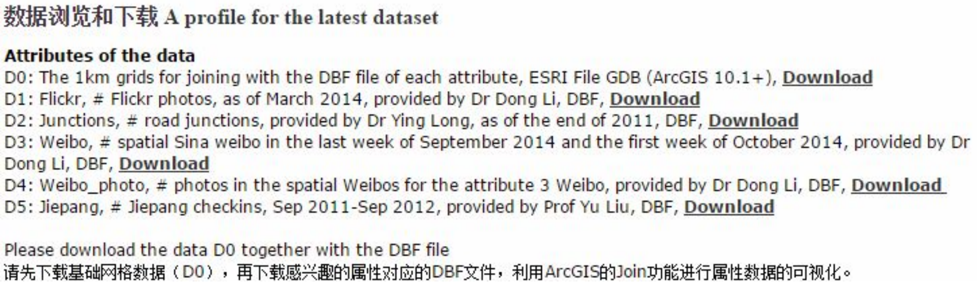
Among all the challenges for community-based open urban data in China, the two most fundamental challenges are: 1) The balance between the benefit of original data holder and that of data user. The major benefit of data user could be realized by making the data accessible and usable. However, guaranteeing the credit of the outcomes based on the data also goes to the original data holder is more important, which would provide motivation for sharing data. Furthermore, the data holders also want their research advantage to be preserved. 2) The gap from accessible to effectively usable. Open data should not only be accessible, but also be effectively usable, which requires high consistency of data format and clear meta-data for recording data specification.

Aiming at solving the above two major challenges to empower urban studies in China, we initiated SinoGrids (Plan Xu Xiake), a crowdsourcing platform for encouraging the standardization (downscaling), sharing and interoperation of micro-scale urban data.

The downscaling is by projecting the original datasets onto a uniform grid. 1km is selected as the current scale for the grid, which is available for both regional analysis and internal urban studies. More scale levels of data products will be added and user will have different level of access according to their own contribution of data. Guidelines and toolsets are freely provided for the standardization (downscaling) of datasets, which, other than being used for data sharing, could also be used as standardization process for multi-source data interoperation. There are many people, who care about and understand cities, needs the support of urban data analysis, but knows little about data processing, e.g. urban planners, government officers, etc. The guideline is written concerning these people. The procedure is explained in detail, intended to lower the technical requirement for data donators, making the platform open to a more general public.

Comparing with other open data platforms, for example, *Data.org* and *ISPIRE*, both of them are government-driven open data infrastructures with detailed data sharing legislation, data specification, etc. As government-driven projects, the main data sources of *Data.gov* and *INSPIRE* are official agencies and data communities, which, in China, still needs time. In China, government-driven platforms are at an early stage. **In this context,** SinoGrids is designed to provide an effective crowdsourcing way for urban researchers, planners and consultants to get urban data in need. As a crowdsourcing platform, the main data sources of SinoGrids are individual data holders, SinoGrids proposes a way to encourage data sharing and improve the usability and interoperability of data. Besides, SinoGrids are mainly focusing on community-generated data, e.g. social media records, instead of government-generated data, providing a public interface for characterizing urban form and urban function instead of the administration perception that *Data.gov* and *INSPIRE* are providing.

In China, *Data Center for Resource and Environment Science* (http://www.resdc.cn/Default.aspx) also provides 1km gridded datasets for China. But those datasets are mainly focusing on the physical side, e.g. vegetation, land use, etc. On the social side, only gridded GDP and population datasets are available. For community-driven data platforms, *Datatang* encourages data sharing by building up tunnels for data trading. SinoGrids is the first to focus on providing open gridded basic urban datasets on the social side to empower urban studies in China.



数据浏览和下载 A profile for the latest dataset

Attributes of the data
D0: The 1km grids for joining with the DBF file of each attribute, ESRI File GDB (ArcGIS 10.1+), **Download**
D1: Flickr, # Flickr photos, as of March 2014, provided by Dr Dong Li, DBF, **Download**
D2: Junctions, # road junctions, provided by Dr Ying Long, as of the end of 2011, DBF, **Download**
D3: Weibo, # spatial Sina weibo in the last week of September 2014 and the first week of October 2014, provided by Dr Dong Li, DBF, **Download**
D4: Weibo_photo, # photos in the spatial Weibos for the attribute 3 Weibo, provided by Dr Dong Li, DBF, **Download**
D5: Jiepang, # Jiepang checkins, Sep 2011-Sep 2012, provided by Prof Yu Liu, DBF, **Download**

Please download the data D0 together with the DBF file
请先下载基础网格数据（D0），再下载感兴趣的属性对应的DBF文件，利用ArcGIS的Join功能进行属性数据的可视化。

Figure 1 Profile for the latest datasets

We designed a scheme to format every dataset characterizing different aspect of urban form or function as a single .dbf file. The users can download selected .dbf files corresponding to the attributes they want, and after joining with the constant 1km grid, the dataset is ready to use. In this way, the users do not have to download datasets with all attributes or download a large grid every time, which dramatically improves the efficiency of data distribution. The independently provided grid is guaranteed constant.

Currently, at the starting stage of SinoGrids, we have received various datasets from generous donators and cooperation intentions from planning institutes and government departments. The available datasets mainly focus on the social aspect of urban behaviors, including social media datasets, road junctions, etc. Shortly, more datasets, like the population grid, the public infrastructure grid, would be added. SinoGrids is a project initiated under *Beijing City Lab*, a virtual community for urban planners and researchers in China with over 40 research fellows, 42 junior research members and over 8000 followers, all of which are potential data donators and users for SinoGrids.

Interactive visualization is generated based on gridded data to make the dataset illustrative to the more general public with no technical background and no interest in data processing.

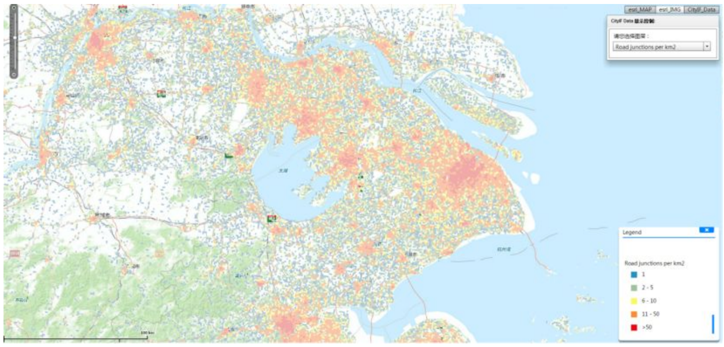


Figure 2 An interaction map of road junctions per square kilometer in Shanghai and nearby area (the Yangtze River Delta)

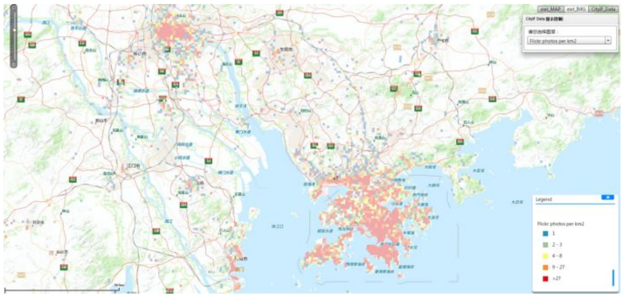


Figure 3 An interaction map of Flickr photos per square kilometer in Hong Kong and nearby area (the Pearl River Delta)

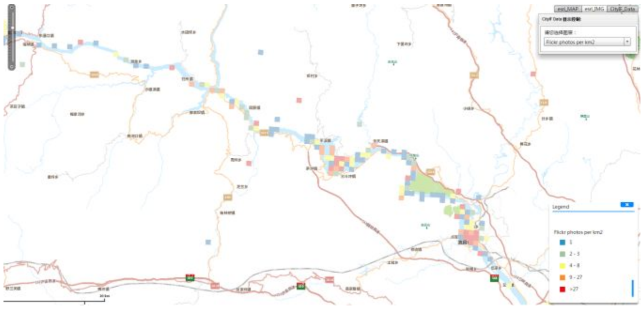


Figure 4 An interaction map of Flickr photos per square kilometer around the Three Gorges and along the Yangtze River (Tourist Attractions)

3.1 User Evaluation on SinoGrids

For user evaluation of SinoGrids, we proposed a formal questionnaire survey in order to collect feedback and comments from previous and potential data donators and data users. A total of 158 effective questionnaires were collected and analyzed. Among the surveyed individuals, 68.99% are urban researchers and students, 31.65% are urban planners and the rest are GIS specialists, government officers, etc. As mentioned before, 79.75% of the surveyed individuals claim that they 'often' meet the circumstance of lacking proper data for urban analysis.

On the side of the original data holder, the main concerns preventing them from opening data are: 1) The possibility of losing research advantage (50%); 2) Constraints on data distribution proposed by the institute or original data generator (53%);3) Extra workloads (35%); 4) Lack of real credit for data holder (54%). SinoGrids is intended to help relieve the major concerns and promote crowdsourcing urban data opening. The survey shows that 91.51% of the surveyed individuals "have fewer concerns donating data" through the SinoGrids way. The feedback shows that the downscaling process does contribute to solving the benefit conflict between data holder and user, making the original data holders more willing to share the data. In addition, we provide data holders with the detailed manual, guideline and GIS tool for aggregating micro-scale data onto the grid. Thus, the data re-generating process is simplified and smoothed, which requires little data processing experience and could be widely used. The questionnaire survey also shows that 94.94% of the surveyed individuals think the SinoGrids way is helpful in improving the efficiency and effectiveness of the interoperation of urban data from various sources.

On the other side of data users, we surveyed the current data sources for urban analytics. Majority of the surveyed individuals obtained urban data (69.62%) from project collaboration, 50.63% used commercial data services, 37.34% got data through online data sprawl and only and 31.65% has benefited from online open datasets. Datasets from SinoGrids has been downloaded and effectively used by urban planners, researchers and consultants. We interviewed several users who have used datasets from SinoGrids in their urban application. We gathered feedbacks, like "finally found free data for social media on SinoGrids, though downscaled, but good enough for my research" and "provides an effective way for data interoperation".

There are also debates on SinoGrids as well. Some claim that there are drawbacks on regular grids for they could never match real world features. Thus, it is better to use units in irregular polygons, e.g. administration blocks. However, blocks vary in size, and China is a rapidly changing country, it is not possible to find a set of high-resolution blocks that is applicable everywhere. Regularly gridded datasets are more general, simple and more proper for data sharing. When needed, high-resolution uniformly gridded data could also be further aggregated to irregular units, e.g., district, block, etc.

3.2 Digital Desert: A son project of SinoGrids

What can we do with datasets in SinoGrids? Here's a son project of SinoGrids, Digital Desert, targeting at analyzing the areas where social media data hardly covers. Study results based on social media data in these areas are less reliable. Urban data analysis does not necessarily provide equally valid results everywhere. Digital deserts could further support the estimation of error for analysis results based on social media data.

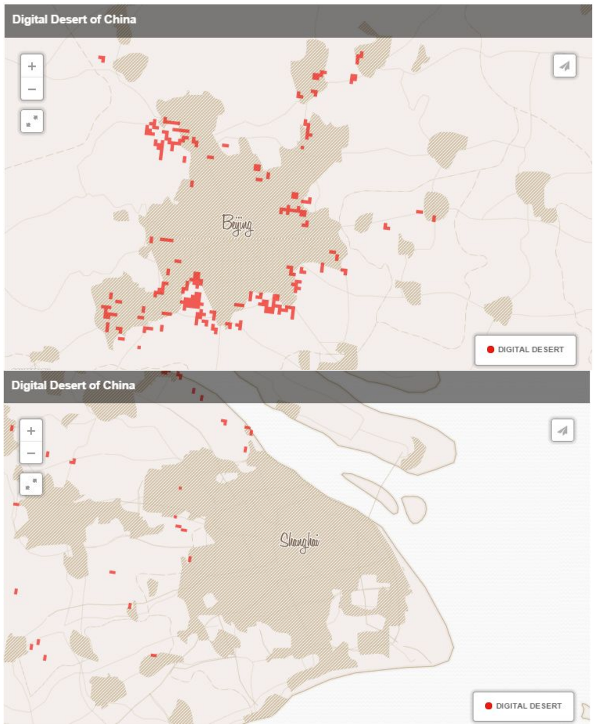
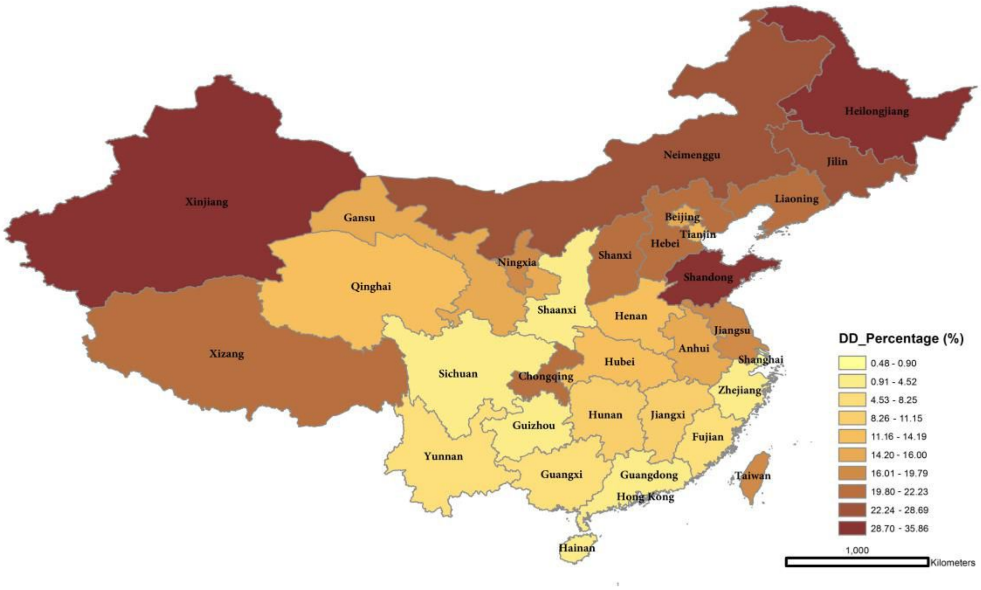


Figure 5 Interactive maps of digital deserts in and around Beijing, Shanghai and their nearby area.

In this evaluation of digital deserts, social media data from different sources, Flickr and Weibo, are considered. The grids with a total number of social media data records less than a certain threshold value in this case, 6, are considered to be digital desserts, where social media data could hardly characterize urban behaviors due to limited amount of data. The threshold is determined by local knowledge and judgment, as well as various experiments.



6

Figure 6 Degree of Digital Desert proportion at province level

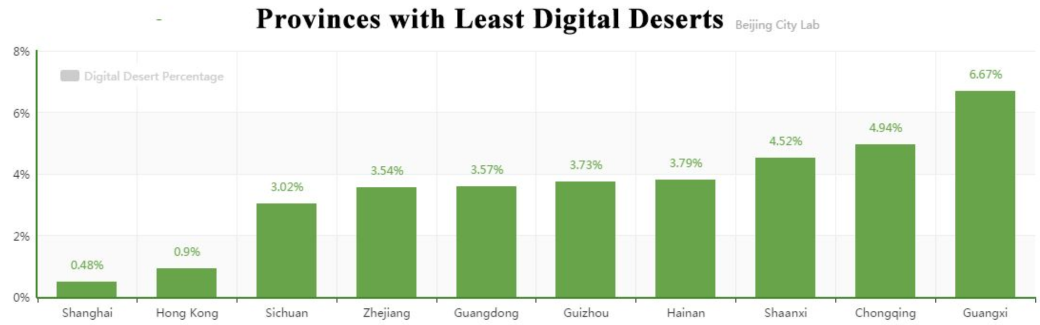


Figure 7 Ten provinces with least Digital Deserts

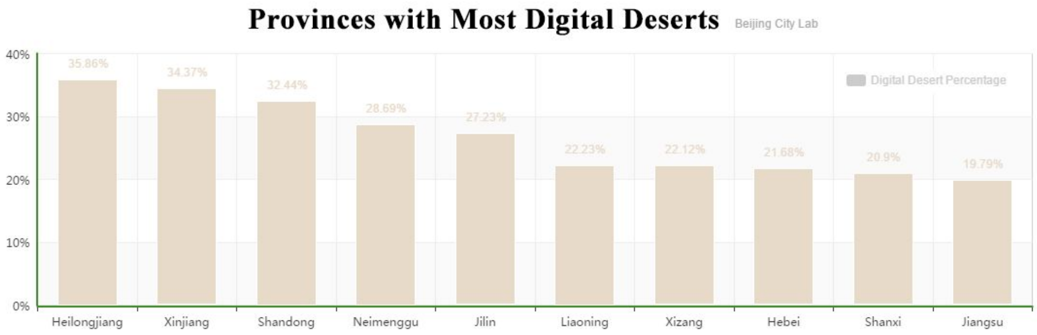


Figure 8 Ten provinces with most Digital Deserts

The degree of Digital Desert proportion at province level and city level, respectively, are calculated and visualized based on SinoGrids gridded social media datasets. Furthermore, the ranking of 10 provinces with most/least digital desert proportion and 15 cities with most digital desert proportion are calculated and made open online.

At provincial level, it is identified that more-developed provinces (or SAR), Shanghai, Hong Kong, Sichuan, Zhejiang and Guangdong has the lowest proportion of digital desert, while Heilongjiang, Xinjiang, Shandong, Inner Mongolia, and Jilin has the highest proportion of urban land without noticeable social media coverage.

At municipal level, there are totally 77 cities with the percentage of digital desert under 2%. Among these cities, 14 are directly-governed cities, provincial capitals and sub-provincial level cities. Due to larger population and density, better economic development, better access to Internet and fast-paced lifestyle, it is reasonable that more-developed cities, like the aforementioned cities, have a relatively lower percentage of digital desert. The theory should also be applicable for most cities in Beijing-Tianjin Area, Yangtze River Delta and Pearl River Delta, the three most productive and wealthy areas in China. But this is not always true. There are also major cities with high digital desert rates, for example, Beijing (15.31%), capital city of China. Beijing has the largest area of urban construction land in China, and it also has the largest net area of digital desert, 385 km^2. Most of the digital deserts lie near the rural-urban fringe, and in between Beijing and its satellite towns. Having so much digital desert for a metropolitan is mainly due to rapid urban development style. Cities in China, especially major cities are growing fast with dramatically expanding urban boundaries. The newly constructed urban area needs time for people to move in, for the improvement of urban infrastructures, etc. before it becomes real urban area. In the case of Beijing, the occurrence of massive digital desert is also due to its dramatic urban sprawl mode. In the contrast, 49 of the 77 cities with low percentage of digital desert (< 2%) are small cities with less than 30 km^2 of urban construction land, due to more stable urban expanding mode or natural physical constraints. Last but not least, even as minor cities, tourist destinations, e.g. Lijiang, Zhangjiajie, etc. has high social media coverage, for the population density is larger and people use social media more for sharing and memorizing while traveling.

Table 1. Examples of Cities with low percentage (<2%) of digital desert (PDD).

City Name	PDD(%)	Possible Cause
Shanghai	0.39	Directly Governed
Chengdu	1.57	Provincial Capital
Guangzhou	1.52	Provincial Capital
Ningbo	1.73	Sub-Provincial
Xiamen	1.80	Sub-Provincial
Foshan	1.93	Third Largest City in Guangdong
Mianyang	0.00	Second Largest City in Sichuan
Huzhou	0.00	Historical City
Sanya	0.00	Tourist Destination, Phisical Constraint
Lijiang	0.00	Tourist Destination, Minor City

Zhoushan	0.00	Minor City, Physical Constraint
Yan’an	0.00	Historical City, Minor City

The nationwide distributions of digital desert proportion, the rankings, together with the extracted grids identified as digital deserts, provides unique and interesting facts and reference to researchers, planners, consultants and the general public based on SinoGrids.

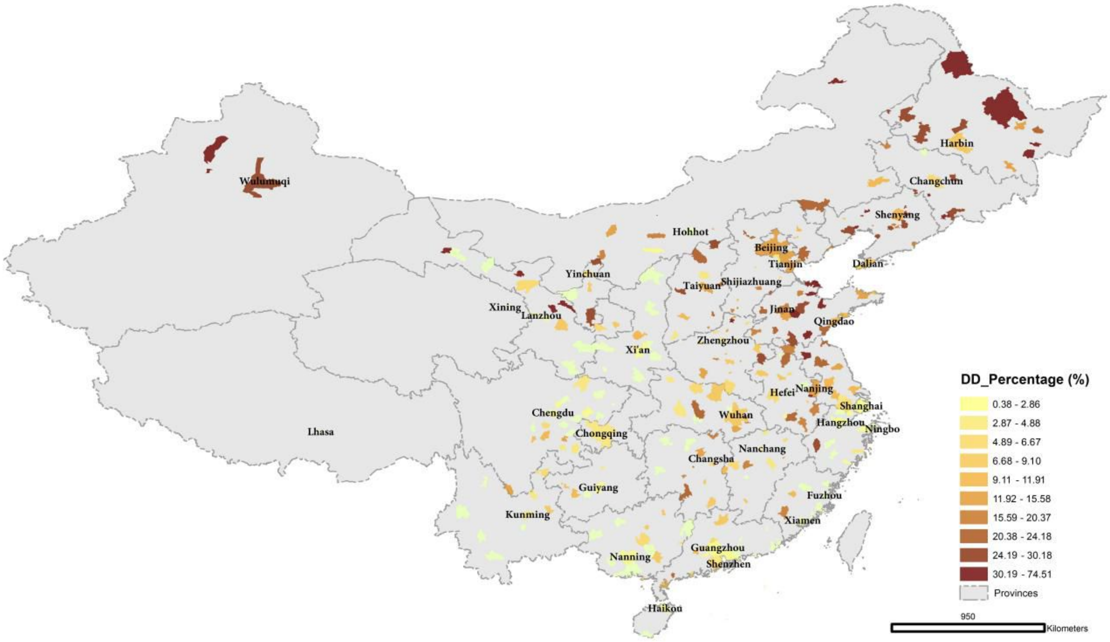


Figure 9 The degree of digital desert proportion at city level for China’s major cities

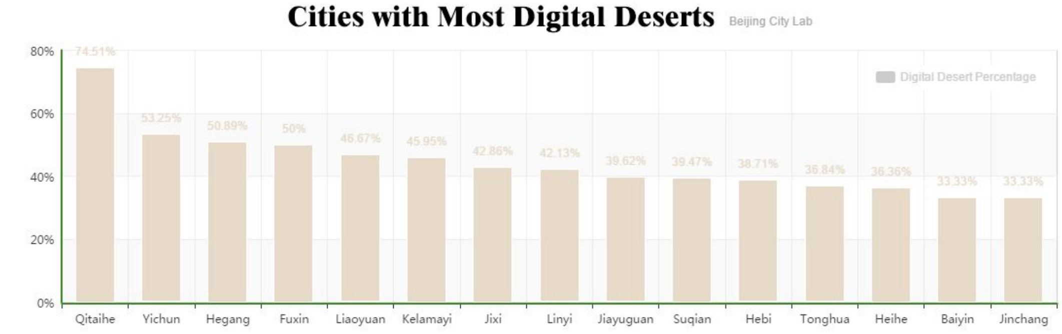


Figure 10 Fifteen cities with most Digital desert

Shortly, more social media datasets, e.g. Jiepang, are to be taken into account to make the calculated digital desert more reliable. The identification of digital deserts would be more valid as more social media datasets are donated, standardized, made public on SinoGrids and got involved in digital desert detection.

4 Concluding remarks

In this short paper, we discussed the current situation of open urban data in the whole World and China alone, and especially listed the challenges that open urban data in China are facing, among which, two major challenges are: 1) The balance between the benefit of original data holder and that of data user; 2) The gap from accessible data to effectively usable data.

SinoGrids, as a crowdsourcing sharing platform for basic urban data in China, is aiming at solving the challenges mentioned above. The main approach is through a uniform-grid-based downscaling standardization process. The benefits are obvious. 1) High-resolution data (e.g. geo-tagged social media points) from the data holder are downscaled onto a uniform grid so that the advantage of proposing further research is preserved for the data holder. Meanwhile, the data user also benefits from having more usable dataset to empower regional analysis and inter-city studies. 2) The standardization (downscaling) process is also enabling further data interoperation by normalizing data format and spatial analysis unit. Guidelines and toolsets are freely provided. The guidelines are written in details, making SinoGrids more efficient and applicable to the general public, including urban planners and consultants with little data processing experience. Furthermore, interactive visualizations of gridded data are made for providing readable data for the general public that has no resource and interest in quantitative data analysis.

A human participated test is proposed for user performance evaluation of SinoGrids. The survey shows that lacking proper data for urban data application has been an often-met situation. But still few people has an open mind in sharing data with others. Among the main concerns of data holders is the the possibility of losing data advantage. SinoGrids has contributed to relieve the concerns. The proposed survey shows that 91.51% of the surveyed individuals “have fewer concerns donating data” via the SinoGrids way. The user evaluation also shows that 94.94% of the surveyed individuals think the SinoGrids way is helpful in improving the efficiency and effectiveness of the interoperation of urban data from various sources.

A son project of SinoGrids, Digital Desert, as a typical application of SinoGrids, is proposed for the purpose of identifying areas with little social media data coverage using gridded social media datasets from SinoGrids. Study results in these areas are not reliable due to insufficient data. The distribution of digital data proportion at province and city level are illustrated respectively. Rankings of 10 provinces with most/least digital deserts and 15 cities with most digital deserts, together with example cities with least digital deserts are generated. The nationwide distributions of digital desert proportion, the rankings, together with the extracted grids identified as digital deserts, provides unique and interesting facts to researchers, planners, consultants and the general public based on SinoGrids. Further research could be done looking into the underlying patterns of the generation and sprawling of digital deserts. And the calculation of digital deserts would be more reliable as more social media datasets are donated, standardized and made public on SinoGrids. The project of digital desert is also a solid example that proves SinoGrids empowers the production of valuable urban analysis results by making more sharable and usable urban datasets open.

SinoGrids is not perfect. In the future, there are still a lot of things to do. 1) Pairing each dataset with standard meta-data. Data from different sources, in a different format, focusing on different topics may have variations in the acquisition methods, and have different spatial and temporal scale. Also, data have history, datasets may have gone through different processing steps. Besides, researches based on the same datasets may be interested to compare with each other. Thus, it is important to pair each dataset with a "name card", recording the spatial-temporal scale, capturing method, processing history, etc., to make the data more reusable and redistributable and prevent misusage. 2) Donation bonus. In the future plan of SinoGrids, the datasets are classified into two types, 5 km gridded and 1km gridded, and for every user, the availability of more precise datasets could be gained by making more contributions to other data users. Everyone is both data holder and data user. More sharing, more gains. The credit system is designed to encourage data opening. 3) Data citation. For current open data portals, the common way for data citation is that data users are required to cite the name of the data platform. As SinoGrids is a crowdsourcing platform, we plan to set a rule to require citation of the original data donator, giving the credit back to the contributor of the dataset and encouraging further data opening. 4) More datasets. Of course, the number of datasets is small at this early stage of SinoGrids. With more datasets donated, standardized and made public, SinoGrids would be more energetic, and be able to further strengthen the interaction among the urban research, planning and consulting community, by encouraging everyone to empower each other with open data sharing.

Reference

Aoun, N., Matsuda, H., & Sekiyama, M. (2015). Geographical accessibility to healthcare and malnutrition in Rwanda. Social Science & Medicine, 130, 135-145.

Bai, X., Shi, P., & Liu, Y. (2014). Society: Realizing China's urban dream. Nature, 509(7499), 158.

Batty, M. (2013). The new science of cities. Cambridge, UK: The MIT Press.

Batty, M. (2012). Smart cities, big data. Environment and Planning B, 39(2), 191.

Center for International Earth Science Information Network (CIESIN), et al. 2000. Gridded Population of the World (GPW), Version 3. Palisades, NY: CIESIN, Columbia University. Available at http://sedac.ciesin.columbia.edu/gpw/index.jsp, retrieved Mar. 14, 2015.

Center for International Earth Science Information Network (CIESIN), et al. 2000. Global Rural-Urban Mapping Project (GRUMP), Version 1. Palisades, NY: CIESIN, Columbia University. Available at http://sedac.ciesin.columbia.edu/data/collection/grump-v1, retrieved Mar. 14, 2015.

Crooks, A., Pfoser, D., Jenkins, A., Croitoru, A., Stefanidis, A., Smith, D., ... & Lamprianidis, G. (2014). Crowdsourcing urban form and function. International Journal of Geographical Information Science, (ahead-of-print), 1-22.

Enserink, B., & Koppenjan, J. (2007). Public participation in China: sustainable urbanization and governance. Management of Environmental Quality: An International Journal, 18(4), 459-474.

Guo, H. (2014). What is the role for government, when big data comes. Harvard Business Review. Available at: http://www.hbrchina.org/2014-12-02/2623.html?mobile, accessed Mar 22, 2015.

Gurin, J. (2014). Open data now: the secret to hot startups, smart investing, savvy marketing, and fast innovation. McGraw Hill Education.

Gurstein, M. B. (2011). Open data: Empowering the empowered or effective data use for everyone?. First Monday, 16(2).

Haklay, M., & Weber, P. (2008). Openstreetmap: User-generated street maps.Pervasive Computing, IEEE, 7(4), 12-18.

Han, H., Yu, X., & Long, Y. (2015). Discovering functional zones using bus smart card data and points of interest in Beijing. arXiv preprint arXiv:1503.03131.

Huang, B., Wu, B., & Barry, M. (2010). Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. International Journal of Geographical Information Science, 24(3), 383-401.

Kitchin, R. (2014). The real-time city? Big data and smart urbanism. GeoJournal, 79(1), 1-14.

Lerman, K., & Ghosh, R. (2010). Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks. ICWSM, 10, 90-97.

Liu, X., Song, Y., Wu, K., Wang, J., Li, D., & Long, Y. (2015). Understanding urban China with open data. Cities. 47, 53-61.

Long, Y., & Thill, J.-C. (2015). Combining Smart Card Data, Household Travel Survey and Land Use Pattern for Identifying Housing-Jobs Relationships in Beijing. Computers, Environment and Urban Systems, In press.

Long, Y., Liu, X., Zhou, J., & Gu, Y. (2014). Profiling underprivileged residents with mid-term public transit smartcard data of Beijing. arXiv preprint arXiv:1409.5839.

Open Knowledge (2014). U.S. City Open Data Census. http://us-city.census.okfn.org/, accessed Mar 20, 2015.

Overpeck, J. T., Meehl, G. A., Bony, S., & Easterling, D. R. (2011). Climate data challenges in the 21 st century. Science(Washington), 331(6018), 700-702.

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. Science, 331(6018).

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010, April). Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web (pp. 851-860). ACM.

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., & Deng, X. (2013, August). Exploiting Topic based Twitter Sentiment for Stock Prediction. In ACL (2) (pp. 24-29).

Tranos, E., & Nijkamp, P. (2015). Mobile phone usage in complex urban systems: a space–time, aggregated human activity study. Journal of Geographical Systems, 17(2), 157-185.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. ICWSM, 10, 178-185.

Wilbanks, J. (2014). Open Data. Privacy, Big Data, and the Public Good: Frameworks for Engagement, 234.

World Health Organization. Global Health Observatory (GHO) data. http://www.who.int/gho/urban_health/situation_trends/urban_population_growth_text/en/, accessed Mar.13, 2015.