Cities 47 (2015) 53–61

Contents lists available at ScienceDirect

Cities

journal homepage: www.elsevier.com/locate/cities

Understanding urban China with open data

Xingjian Liu [a], Yan Song [b], Kang Wu [c], Jianghao Wang [d], Dong Li [e], Ying Long [f,*]

[a] The University of Hong Kong, China
[b] University of North Carolina – Chapel Hill, USA
[c] Capital University of Economics and Business, China
[d] Institute of Geographic Sciences & Natural Resources Research, Chinese Academy of Sciences, China
[e] China Academy of Urban Planning and Design, China
[f] Beijing Institute of City Planning, China

ARTICLE INFO

Article history:
Available online 29 March 2015

Keywords:
China
Open data
Urban analytics

ABSTRACT

A solid understanding of urbanizing China – the world's largest and most rapidly transforming urban society – calls for improved urban data provision and analysis. This paper therefore looks at major technological, social–cultural, and institutional challenges of understanding urban China with open data, and showcases our attempt at understanding Chinese cities with open urban data. Through our showcases, we hope to demonstrate the usefulness of open urban data in (1) mapping urbanization in China with a finer spatiotemporal scales; (2) reflecting social and environmental dimensions of urbanization; and (3) visualizing urban China at multiple scales.

Introduction

With more than 700 million urban residents and another 30 million people being added to its cities every year, China becomes the world's largest and most rapidly transforming urban society (Bai, Shi, & Liu, 2014). Urbanization in China has therefore attracted wide attentions from researchers, policy makers, investors, and the general public. The number of articles featuring "urban China" has risen by 600% in the last decade, based on bibliographic information gathered from the ISI Web of Science. Public awareness about a wide range of urban issues, ranging from public safety, air pollution, to food security, has reached a historical height (Enserink & Koppenjan, 2007). Meanwhile, the recently released "National New-Type Urbanization Plan (2014–2020)" has marked a new consensus among the country's ruling elites to foster sustainable and people-centered urbanization patterns (Bai et al., 2014; Wallace, 2014).

A solid understanding of urbanizing China hinges on improved measurements of Chinese cities, although gaining more data is seldom a solution by itself, as rigorous analyses and insightful communications are necessary to deliver derived information to a wide variety of audiences. Recent studies in urban analytics identify open urban data as a promising complement to conventional authoritative data sources in understanding urban form and functions (Crooks et al., in press; Goetz & Zipf, 2012; Jia & Jiang, 2010). Furthermore, for most developing countries that lack conventional data infrastructures, marshalling existing open data from various sources seems to be a viable option. The purpose of this report is therefore twofold. First, it reviews the major technological, socio-cultural, and institutional challenges of assembling openly accessible urban data for Chinese cities. Second, the paper showcases one specific attempt – the Beijing City Lab – at understanding Chinese cities with open urban data.

What are open urban data?

Before discussing the challenges in assembling and using open urban data in China, we briefly provide a working definition of open urban data within the context of our discussion. We use urban data as an overarching umbrella term for all datasets that would characterize some aspects of urban form and functions (Crooks et al., in press). Our definition of open urban data subsequently is therefore broad and refers to urban data that are openly accessible to the public. Preferably such data are available in convenient and adjustable formats, but as we will elaborate, this is not always the case. Open urban data mainly come from three overlapping though different sources (Gurin, 2014): official data portals, big data initiatives, and the broader open data community.

Urban data have conventionally been produced and maintained by government agencies. These traditional urban datasets range from cadastral maps prepared by land departments and traffic

* Corresponding author at: No. 60, South Lishi Rd, Beijing 100045, China. Tel.: +86 10 8807 3660.
E-mail address: longying1980@gmail.com (Y. Long).

http://dx.doi.org/10.1016/j.cities.2015.03.006
0264-2751/

records maintained by transportation authorities, to business transactions monitored by taxation agencies. Recent open government initiatives have propelled governments to grant public access to a greater amount of authoritative data sources. This is best evidenced by mushrooming data portals for US cities in the wake of The Digital Accountability and Transparency Act of 2014 and other related local legislations. For example, NYC OpenData, the official online data portal of New York City (NYC), provides open access to more than 1300 governmental datasets, some of which are quite big in terms of volume.

Big data enable us to gauge urban dynamics at finer spatial and temporal scales (Kitchin, 2014; Yue et al., 2014). The generation of big data is characterized by large volume, great variety, and high velocity. Typical big data in the urban environment include mobile phone activities, vehicle trajectories, public transit smart card data, business catalogues, as well as other information generated by various smart cities programs (Batty, 2012). These types of data often enable researchers to capture urban dynamics at very fine spatiotemporal scales. Many of these datasets are often made available by private data holders via public–private partnerships and data challenges. For example, the 'Data For Development' challenge (http://www.d4d.orange.com/en/home) organized by the Orange Group released mobile phone usage datasets for African countries Ivory Coast and Senegal.

Volunteer Geographic Information (VGI) and Crowdsourcing (Goodchild, 2007; Crooks et al., in press) are major contributors to urban data within the general open data community. As with other grassroots information campaigns (Howe, 2006), urban data in this category are contributed by the general public and represent a 'bottom-up' data generating process. Notable examples include collaborative VGI mapping platforms (e.g., OpenStreetMap) and location-enabled social media applications (e.g., Foursquare, Twitter, and Flickr).

Open urban data can be further differentiated along three dimensions: (1) Online/offline: direct downloading from the Internet is the preferred access mode of open data, and indeed many examples we have given so far are online sources. Nevertheless, we note that not all open urban data can be accessed online. Offline open urban data include *inter alia* hard-copy government reports, historical paper maps, as well as yet-to-be-digitized library collections. (2) Explicit/implicit: Crooks et al. (in press) emphasize the distinction between explicit and implicit urban data. Many open urban data are of the former type, being readily available for use in the desired form. For example, datasets from the NYC OpenData portal are stored in standardized tabular and geospatial data formats. By contrast, implicit urban data capture information about urban dynamics, but cannot be applied directly in analyses. Such implicit data need to be derived and pre-processed. For example, the business cataloguing website Yelp contains a wealth of business establishments (e.g., location, ratings, and descriptions), however such information needs to be crawled from individual webpages and geocoded before any subsequent analyses. (3) Usable/re-usable/redistributable: Open datasets are subject to different sharing requirements and codes. The more open the data, the less strings attached: At one end of the spectrum, open data allow for free use, re-use, and re-distribution by anyone. However, on other occasions, the use of datasets may be restricted for education purposes. For example, datasets released by the aforementioned 'Data for Development' data challenge are for education and research only.

Open urban data in China

The increasing supply of scientific data, spearheaded by numerous open data initiatives, has transformed research practices in many fields (see for example, a *Science* magazine special issue on data and new science initiatives; *Science* Vol. 331, Issue 6018). These new waves of data initiatives are also opening up new research opportunities in urban studies and planning (Batty, 2013; Crooks et al., in press). Meanwhile, researchers still lament the dearth of urban data in developing countries. While European and North American cities are being measured with greater precision and granularity, cities in the developing world are oftentimes less well-charted. In others words, there seems to be a widening gap in terms of openly available data between cities in developed and developing worlds (Graham, Hogan, Straumann, & Medhat, 2014; Lang, 2011). Overcoming such gaps in data availability seems to be of high priority for urban studies and planning in developing countries in general and China in particular.

As in many developing countries, government agencies are the primary supplier of urban data in China and exercise tight data regulations for reasons such as national security. Nevertheless, the Chinese government is pushing for more open access to data via open government initiatives and loosening control of data use (see for example, the launch of an official national data repository: http://data.stats.gov.cn/). Recasting Reichman, Jones, and Schildhauer (2011)'s insights and framework about challenges of open data in ecology, we have identified the following technological, social–cultural, and institutional challenges of collecting open urban data in China.

A first technological challenge concerns the dispersion of urban data, similar to the cases in ecology (Reichman et al., 2011). In general, efforts to produce urban data are rather fragmented, as most publicly available data about Chinese cities are scattered in online and offline sources. As discussed previously, data sources range from user-generated contents on the Internet, governmental yearbooks and websites, datasets released by private companies,[1] to datasets produced by research projects. Additionally, there exist substantial overlaps between data collected via different venues, creating discrepancies for data use and wasting resources. For example, due to the rather complicated *tiao/kuai* administrative framework in China (Mertha, 2005), (urban) land use data are being collected independently by a number of government agencies, such as land use and resources, urban planning, as well as agricultural and forestry departments.

With the dispersion of data sources comes the second technological challenge: the heterogeneity of data. Data from different sources are usually produced and maintained in different formats, focusing on different urban issues, capturing different years, and covering different geographic areas (Reichman et al., 2011). For example, information about urban construction permits is usually available from individual municipal governments' websites (see for example, http://www.bjjs.gov.cn/tabid/660/Default.aspx). Such information is critical for understanding urban growth patterns, as construction permits data capture the 'where' of urban development at a very fine spatial scale. However, permits data of different cities do not come along in the same format: some exist on fully-fledged websites that enable online information retrieval, others take the form of offline archives and require digitization; some are explicit and encoded in tabular forms, others are implicit and need to be derived from plain texts; some are spatially explicit and contain longitudinal and latitude coordinates and are thus readily available for mapping, others need further geocoding.

Furthermore, the pace of urbanization in China posits a unique challenge: it is unprecedented and faster than the frequency of many regular governmental monitoring and census programs.

[1] In addition to businesses operated by large consulting and data companies, proprietary data, ranging from digitized yearbook to sales records of individual companies, can be purchased via many small vendors on taobao.com – the Chinese equivalent of eBay.