Discovering and Accessing Sub-national Statistics and Geospatial Data of East Asian Countries: Trends and Obstacles

by Jungwon Yang[1]

Abstract

The increasing use of geographic information systems (GIS), combined with the wider availability of sub-national statistics, has recently opened up new possibilities for more interdisciplinary academic research in social sciences. Social science researchers have become more and more interested in combining geographic analysis, traditional quantitative and statistical methods in order to test hypotheses and present arguments in more effective ways. However, discovering, accessing, and using the international geospatial data and statistics is still a challenge for the researchers. As the Organisation for Economic Co-operation and Development (OECD) Global Science Forum report (2013) noted, information about the existence of micro-data and the availability for the re-use is often difficult to find. The language barriers, as well as, legal, cultural, and technological obstacles often exacerbate the difficulties of re-using the discovered data. In this paper, I will review types of geospatial data and sub-national statistics of East Asian countries have recently developed via central and local governments, and academic institutions. Also, I will discuss obstacles researchers encountered while using such data in their research.

adoption of data citation and in the promotion of data sharing and its benefits.

Keywords: : GIS, sub-national statistics, East Asia, geospatial data, China, Korea, Japan

Introduction

The increasing use of geographic information systems (GIS), combined with the wider availability of sub-national statistics, has recently opened up new possibilities for more interdisciplinary academic research in the social sciences. Social science researchers have become increasingly interested in combining geographic analysis and traditional quantitative and statistical methods to test hypotheses and present arguments in more effective ways. Researchers are more likely to be interested in interdisciplinary studies as visualizing data can facilitate interpretation and understanding of social scientists' results. People often have some difficulty to interpret the results of quantitative analyses if they do not have a solid grasp of statistical analysis, such as the p value, R square, or confidential levels. Using GIS technology and geospatial data in interdisciplinary research helps people to understand their research even if they do not have comprehensive knowledge of the research methods.

Moreover, the share of the United States' local governments adopting GIS rose steadily from 20 percent in 1990 to nearly 88 percent in 1997. As of 2005, more the 60 percent of municipal web sites had begun to provide interactive GIS features (Ganapati, 2011). A wide variety of population and housing census data is freely available in various data formats and for a range of geographies and time periods from the United States Census Bureau website . The University of Minnesota's National Historical GIS Project is processing and making freely available U.S. census boundaries in GIS format as well as aggregate census data from 1790 to 2012 (*https://www.nhgis.org*). Given the increased availability of geospatial and sub-national statistics, social science researchers can test their hypotheses and explain them in more effective ways. For example, Sinclair et al (2011) uses GIS-coded flood-depth and census data to examine the voting behavior of registered voters in New Orleans before and after Hurricane Katrina. To measure the quality of the urban environment around public housing buildings in Montreal, Apparcicio et al (2008) uses multiple years of census data and geospatial data, such as the 2001 Montreal urban community land use map, a Landstat TM 7 image, the Geobase, and the Quebec topographical database. Researchers can also create their own geospatial data to use in their research, potentially combining it with existing geospatial or other data. Neckerman (2009) uses data both from GIS measures and field observation in New York City to identify disparities in neighborhood conditions, by aggregating geospatial data that represents low-income neighborhoods.

Even though interdisciplinary studies using this GIS technology and geospatial data are prevalent in the western countries, such as the United States, Canada, and some European countries, the discovery, accession, and use of international geospatial data and statistics is still a challenge for researchers. Part of the reason is noted in the OECD Global Science Forum report (2013): "information about the existence of micro-data and availability for re-use is often difficult to find". Moreover, the language barrier, legal, cultural, and technological obstacles often exacerbate reusing the data.

In this paper, I will review the kinds of geospatial and sub-national statistics on China, Korea, and Japan that have been recently developed by the central and local governments, and academic institutions. I will also address what obstacles researchers have encountered in using the data in their research.

Geospatial Data and Sub-national Statistics in East Asian Countries

Macro-level economic and population data of East Asian countries have been collected by the intergovernmental organizations, such as the United Nations Statistics Division, the International Monetary Fund (IMF), and the organization for Economic Co-operation and Development (OECD). The Korean and Japanese central and local governments tend to collect and distribute sub-national statistics, including census, housing and economic data, and industrial enterprise data, as well as geospatial data, such as hydrology, elevation and land cover data. Most statistics collected by the Chinese government, however, are not available from the government website.

China

The primary department of the Chinese government that collects sub-national statistics is the National Bureau of Statistics of China (NBS) . The NBS organizes collection of census data, which is held every ten years. It also collects, processes, and tabulates basic economic, social, and industrial enterprise statistics. Most of the data collected by the NBS, however, is not open to public. On the Chinese version of the NBS website , only brief summaries of the statistical yearbook for the prior three years are available. The Chinese government does not collect geospatial data. Sub-national statistics and geospatial data for China is available from some fee based databases, such as the China Knowledge Resource Integrated (CNKI) , and the China Data Center .

The CNKI database provides China's statistical yearbook (1981-2013), the Chinese Health statistics yearbook (2003-2012), the China Financial yearbook (1949-2013), as well as provincial statistical yearbooks. Yet, the datasets are written in Chinese only. The CNKI database does not have geospatial data for the region of China. English versions of statistical and geographic data for China have been collected by the China Data Center at the University of Michigan. The China Data Center's databases consist of two parts: China Data Online and China Geo-explorer. China Data Online provides the census data (2000 and 2005) on provincial, county, and township levels, provincial and city economic statistics, and yearly and monthly industrial data. China Geo-Explorer , which is the web-based spatial data service of the China Data Center, aggregates government statistics, such as census, economic, and industrial data in a spatially integrated system. This spatial data service supports the creation and export of thematic maps based on built-in data. Unlike ArcGIS software, however, users cannot combine the data which they create or acquire from other resources with the data of China Geo-Explorer. From Geo-Explorer II, users can export geospatial data and statistics in the database as a shapefile for use in GIS software.

Several projects from the U.S. academic sphere aim to build historical geospatial data. For example, the University of Washington's China in Time and Space (CITAS) data sets provides vectorized county level base maps of China and georeferenced socioeconomic data for the period 1982-1993. The CITAS data is currently archived in the Socioeconomic Data and Applications Center (SEDAC) 's China Dimensions data collection . The SEDAC provides citations, copyright and privacy policy information to users. Even though the original China Geo-Explorer database is a fee-based database, the Sidney Gamble Photo digital collection , created by Duke University and applied to basic China Geo-Explorer software, is open to the public. The database currently features photographs dating between 1917 and 1932. The China Historical GIS website of the Harvard Yenching Institute offers datasets such as time series geospatial data called CHGIS13 (221BC -1911CE), the 1820 data which includes spatial data of the Qing Dynasty territory for the year 1820, the ChinaW Data (1820-1893) created by the UC Davis Regional Systems Analysis Project , and the 1911 data which contains spatial data for the provinces of Anhui, Fujian, Gansu, Guangdong, Hebei, Henan, Hubei, Hunan, Jiangsu, Shaanxi, Shandong, Zhejiang, and Zhili for the year 1911. The CHGIS website also contains the 1990 CITAS data, the 1990

GNS place names, the 1997 CITAS provinces data, topographic images and raster data derived from GTOPO-30 Digital Elevation Model data, and other supplemental datasets.

As Thompson (2010) notes, non-governmental organizations tend to collect sub-national data targeted toward a particular context or issue. For example, the Harvard Yenching Institute focuses on collecting geospatial data for the pre-modern period of China. The CITAS is more likely to focus on collecting and providing relatively current geospatial data. Another interesting finding is that academic institutes which provide open geospatial data are more likely to archive their data into other bigger and stabilized academic institutions. For example, the CITAS data, aggregated by the University of Washington, is currently deposited in the SEDAC and the CHGIS datasets. As a result, even if an academic organization is no longer able to provide their data to users, the original data, metadata, and citation information are available from other sources.

Korea

Korean central and local governments produce many sub-national statistics, including population, household, employment, prices, health, environment, agriculture, mining, energy, transportation, business, and education data, which are freely accessible from the Korean Statistical Information Service (KOSIS) web site . In addition to domestic statistics, the KOSIS website also provides international statistics compiled by the IMF, OECD and UN.

The English version of the KOSIS website offers statistics at the provincial level, including special self-governing provinces (teukbyeoljachi-do), special cities (teukbyeol-si), and metropolitan cities (gwangyeok-si). However, municipal level data , such as cities (si) , counties (gun), districts (gu), towns (eup), townships (myeon), neighborhoods (dong), and villages (ri), are only available from the Korean version KOSIS website.

In the case of geospatial data of Korea, the National Geographic Information Institute (NGII) of Korea provides aerial photos (1966-2012), satellite images (1973 -2004), orthophotos (2005-2011), and DEM files (2005-2009) to users for a small fee . Aerial photos from the 1940s and 1950s are freely available to the public. The standard image format is National Image Exchange (NIX), which contains the image files compressed by JPEG 2000 as well as associated metadata. Based on the information disclosure act of Korea , all Korean citizens, state agencies, local governments, government-funded institutions, and public authorities, as prescribed by presidential decree (for example schools, construction companies, non-profit corporations related to social welfare), can request to access official government documents (including electronic documents), drawings, photographs, films, tapes, slides and other similar recorded mediums. Access to NGII's data from a foreign IP address is strictly prohibited , so researchers cannot acquire the NGII data from overseas.

The majority of provincial governments (gun) and metropolitan cities in Korea provide GIS information for their regions. For example, the Seoul Metropolitan Government provides a city information map service in Korean and English languages. The English version of the map service is an interactive online map which contains basic information for tourists, such as government office building locations, hospitals, schools, transportation information, and historic sites. The Korean version of the interactive mapping service for the city of Seoul additionally provides administrative boundaries, buildings, transportation, facilities related to welfare services, environment, land use and road information. Users can overlay multiple kinds of geospatial data onto a map and freely download the customized map to as an image file. The city of Seoul also offers open API service to Korean citizens but based on Article 21 of the Public Survey and Overseas Export Ban Act, international map mashup services are strictly prohibited .

The map service web sites of Incheon city , Gyeonggi , South Gyeongsang , and Gangwon provinces also offer customized map services for overlay of aerial photos, a base map, transportation and other types of statistical information. The customized map can be downloaded to as an image file without fee. Other provincial governments, such as Jeju and South Chungcheong provinces and metropolitan cities, such as Busan, Daegu, Gwangju, and Daejeon cities, also provide a GIS map service, but it requires the up-to-dated Internet Explorer browser to work. It is quite cumbersome for other internet browser users, such as Chrome and FireFox to access their data. Moreover, even if users use the Internet Explorer browser, they cannot access the data if they use out-of-dated Internet Explorer program.

In sum, the Korean central and local governments have a strong interest in developing and distributing geospatial data as well as sub-national statistics to the public. However, the GIS services are mainly provided to serve Korean citizens. Overseas researchers often encounter difficulty to acquire and use the geospatial data.

Japan

The Statistics Bureau and Ministry of Internal Affairs and Communication of Japan collect both sub-national statistics and geocoded census data, and distribute via the Statistics Bureau website . Current national level statistics can be found in the Japan Statistical Yearbook series section of the website, in both PDF and Excel formats. Historical statistics from 1868 to 2011 are available from the Historical Statistics of Japan section of the website. All the data from the website is available in Excel format. The current sub-national statistics report called Social Indicators by prefecture 2014   can be accessed via the Social Indicators by Prefecture section of the web site. It contains 608 social indicators and 571 items of basic data for the sub-national areas . Time series data of sub-national statistics is available from e-Stat , the official statistics site of Japan. The time series data for prefectures, however, are not accessible on the English version of the site. Geocoded census and socio-economic data also can be downloaded from the e-Stat web site . Currently, geocoded census data (2000, 2005, and 2010), the establishment and enterprise census (2001 and 2006), the economic census data (2009), and the census of agriculture and forestry data (2005 and 2010) can be downloaded in shapefile format. The Japanese version the e-Stat website also offers an interactive map service, where a user can download customized images.

The Statistics Bureau of Japan states that their data can be used for "compiling social indicators, by research institutes, universities and colleges for regional characteristic analysis, modeling to analyze regional development plans, and modeling to measure administrative performance, as base data for compiling welfare indicators by region, and for investigating and improving social statistics" . In other words, there is no restriction for accessing the data from overseas. Another resource of geospatial data is the Global Map Japan database of the Geospatial Information

Authority of Japan (GSI) which is part of the Ministry of Land, Infrastructure, and Transport of Tourism of Japan. This database provides the land cover, vegetation, administrative boundary, population, and transportation data for the year 2000, 2006 and 2011 without fee. They also do not prohibit the access from overseas. Based on the copyright laws of Japan, as well as an international treaty, people can download data without consent from the GSI if they use the data in small quantities for non-commercial purposes.

Historic geospatial data for Japan is available from Harvard University's Japan Data Archive . Currently, administrative boundaries for the Tokugawa and Meiji periods are available in shapefile format. These data are also available from the GeoData@Tufts database . In addition, elevation data for 1996 and administrative boundaries for the 1990s can be found on the website. All of Japan Data Archive's data is open to public; however, citation and copyright information is not available on the website.

In sum, the collection and distribution procedures of government data in Japan is highly centralized. They are more willing to share their geospatial data with overseas users, as compared to the Korean and Chinese governments.

Trends and Obstacles

Data Collection and Distribution

The central governments in all three East Asian countries tend to collect sub-national statistics. Most of the Chinese sub-national statistics are not available from the government website, but all sub-national statistics for Korea and Japan are freely available from their Statistics Bureau websites. The Chinese government does not collect geospatial data. Most of the available geospatial data for China is therefore produced by overseas academics or institutions. Korea's geospatial data is collected and distributed by the NGII. The local governments of Korea are eager to provide the geospatial data of their reasons as well. Based on the law, the Korean citizens and academic institutions located in Korea can access this data without difficulty, but the data from NGII, the main provider of geospatial data, cannot be accessed from overseas. In the case of Japan, all the sub-national and geospatial data is collected and distributed by the Statistics Bureau of Japan. In addition, the Geospatial Information Authority of Japan (GSI) provides the geographic data. Restriction of access to Japanese geospatial data is quite minimal.

Copyright and Open Access

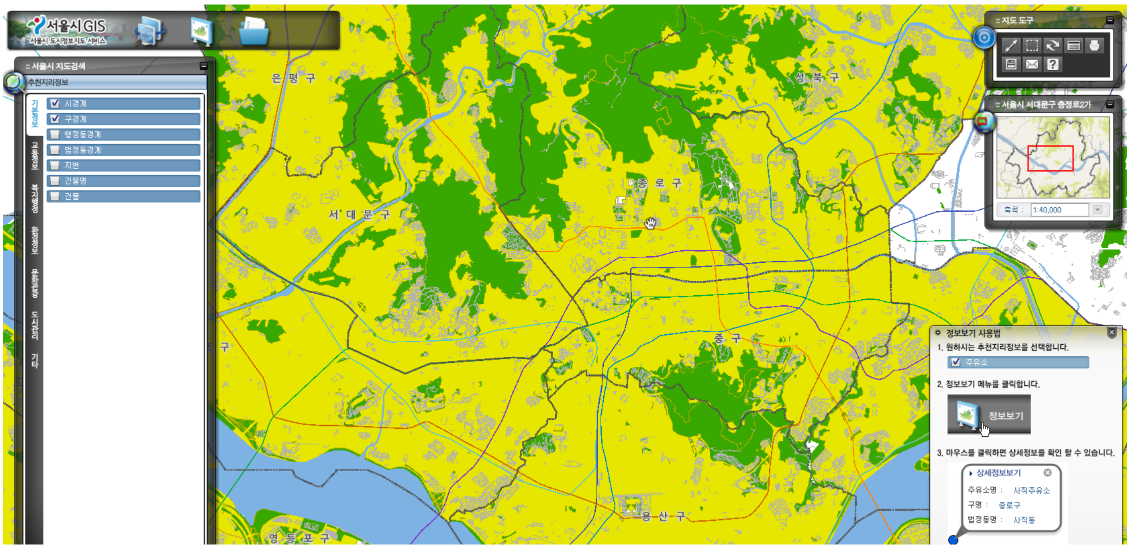
Willingness to support open data access differs across these countries. The Japanese government puts little restriction on either domestic or overseas use of their geospatial data. Most of the Korean statistical and geospatial data created by the central and local governments are open to domestic users. But, the acquisition of geospatial data from overseas is restricted. Both Korean and Japanese governments provide citation information, the data collection method, and copyright information from their websites. Chinese statistics and geospatial data, however, are not open to public. Rather, researchers must subscribe to a fee-based database to access the data. The historic geospatial data, created by academic institutions in the United States, is free to use. Most of the academic institutions provide citation, copyright and disclaimer information for the data.

Reliability of Data

All three countries' governments provide sufficient enough information about data collection procedures. The Korean and Japanese governments also state that they follow the IMF's data classification standard in the process of data collection. Therefore, we assume that the reliability of data created by these governments is quite high. Most of the academic institutions in the United States also provide information about their data collection procedures. Yet, the China Data Center does not provide the sources of data or information regarding data collection procedures. Given these conditions, the question of the reliability and the accuracy of data (economic data in particular) of the China Data Online have long been raised (Chua, 2012) .

Language Barrier

As currently available sub-statistics and geospatial data of China are developed in academic institutions in the United States, the accompanying information, including metadata, is usually written in English. Thus the language barrier to use the Chinese data is relatively low for users, compared to Korean and Japanese data. The majority of sub-national statistics and geospatial data for Korea and Japan are only available in their respective languages. Specifically, menu and the mapping options for the interactive map service website, where users extract customized maps, often cannot be translated into English (see Figure 1 and 2). People who cannot understand Korean and Japanese languages will obviously be unable or have difficulties using these mapping services.



정보보기 사용법

1. 원하시는 추천지리정보를 선택합니다.

☑ 주유소

2. 정보보기 메뉴를 클릭합니다.

정보보기

3. 마우스를 클릭하면 상세정보를 확인 할 수 있습니다.

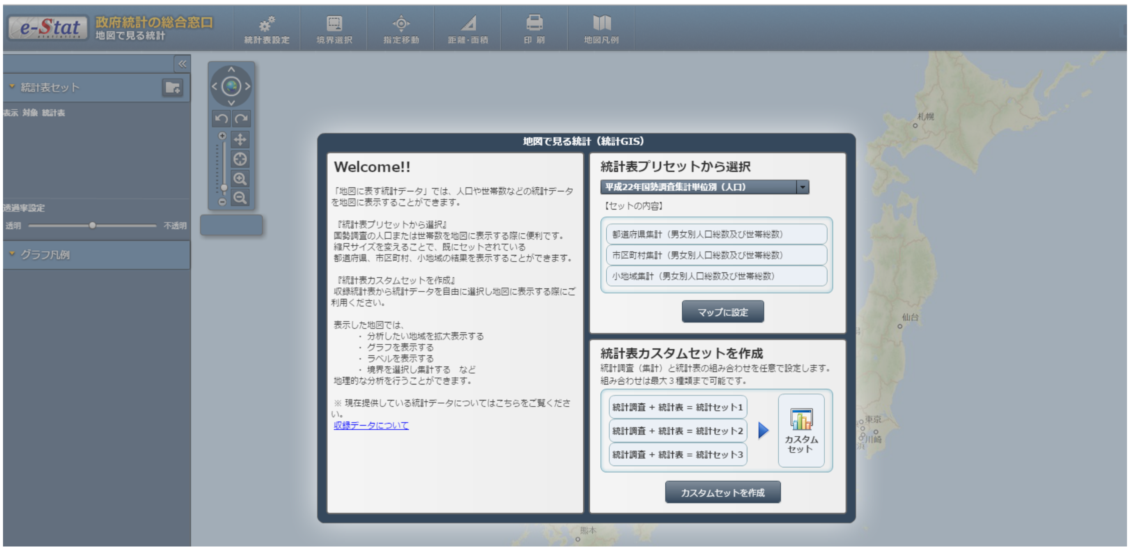
상세정보보기

주유소명 : 사직주유소

구명 : 종로구

법정동명 : 사직동

Figure 1 - Seoul Metropolitan City's GIS portal (Korean Version).



地図で見る統計（統計GIS）

Welcome!!

「地図に表す統計データ」では、人口や世帯数などの統計データを地図に表示することができます。

『統計表プリセットから選択』
国勢調査の人口または世帯数を地図に表示する際に便利です。
縮尺サイズを変えることで、既にセットされている
都道府県、市区町村、小地域の結果を表示することができます。

『統計表カスタムセットを作成』
収録統計表から統計データを自由に選択し地図に表示する際にご利用ください。

表示した地図では、
・分析したい地域を拡大表示する
・グラフを表示する
・ラベルを表示する
・境界を選択し集計する　など
地理的な分析を行うことができます。

※ 現在提供している統計データについてはこちらをご覧ください。
収録データについて

統計表プリセットから選択

平成22年国勢調査集計単位別（人口）

【セットの内容】
都道府県集計（男女別人口総数及び世帯総数）
市区町村集計（男女別人口総数及び世帯総数）
小地域集計（男女別人口総数及び世帯総数）

マップに設定

統計表カスタムセットを作成

統計調査（集計）と統計表の組み合わせを任意で設定します。
組み合わせは最大３種類まで可能です。

統計調査 + 統計表 = 統計セット1
統計調査 + 統計表 = 統計セット2
統計調査 + 統計表 = 統計セット3

カスタムセット

カスタムセットを作成

Figure 2 - E-State GIS portal (Japanese version).

Technical Concern

Some Korean central and local government websites only can be accessed via the Internet Explorer browser. Moreover, even if a user uses the Internet Explorer browser to access the website, the user cannot access the data without installing a current version of Internet Explorer. This technical requirement may add a small burden to overseas researchers' attempts to acquire the data, especially as many United States libraries and computer labs at academic institutions do not allow to users to install new software without administrative access due to security issues. This may be a minor obstacle as many computers come with Internet Explorer installed, or researchers could most likely locate a computer with it installed simply to download the data.

In the case of the China geo-explorer database, its thematic map service does not have an overlay option, so a user can only create a thematic map with a single set of data. Also, census data from 2010 is now available in the China Geo-Explorer database, but the statistical data is not available from the China Data Online database. As a result, researchers have to find an additional database to get the numeric data.

Conclusion

After investigating these data resources, I conclude that we can find quite a large amount of sub-national statistics and geospatial data from government, as well as academic institutions in the United States. However, the unreliability of data (China), the language barrier (Korea and Japan), copyright and legal restrictions (China and Korea), and technical issue (China and Korea), make it difficult for researchers to find and use data for East Asian countries. In the academic library setting, collaboration among specialists and librarians in area studies, copyright, data, GIS, government documents, and maps, will be critical to overcome these legal, cultural, and technical issues and to support social scientists' interdisciplinary studies.

Reference

Apparicio, P., Seguin, A. & Naud, D. (2008) The Quality of the Urban Environment Around Public Housing Building in Montreal: An Objective Approach Based on GIS and Multivariate Statistical Analysis. Soc Indic Res. 86. p. 355-380. DOI 10.1107/s11205-007-9185-4

Bosak, K. & Schroeder, K. (2005) Using Geographic Information Systems (GIS) for Gender and Development. Development in Practice. 15(2). p. 231-237

Chua, H. (2012) Indiastat and China Data Center Online: an evaluation and comparison. Reference Reviews. 26(2).

Ganapati, S. (2011) Uses of Public Participation Geographic Information Systems Applications in E-Government. Public Administration Review, May/June 2011. p. 425-434.

Ghiradeli, A. Quinn, V. & Foerster, S. (2010). Using Geographic Information Systems and Local Food Store Data in California's Low-Income Neighborhoods to Inform Community Initiative and Resources. American Journal of Public Health. 100(11). p. 2156-2162.

Imai, H., Keiko I., Kazuo I., & Koichi K. (2003). GIS Infrastructure in Japan—Developments and Algorithmic Researches. Nontraditional Database Systems. 5. p. 130-145.

Li, Y.(2012). The spatial variation of China's regional inequality in human development. Regional Science Policy & Principle. 4(3). p.263-278

Neckerman, K. et al.(2009). Disparities in Urban Neighborhood Conditions: Evidence from GIS Measures and Field Observation in New York City. Journal of Public Health Policy. 30. p.264-285.

Sinclair, B., Hall ,T., & Alvarez, M. (2011). Flooding the Vote: Hurricane Katrina and Voter Participation in New Orleans, American Politics Research. 39(5). p.921-957. doi:10.1177/1532673X10386709

Thompson, K. (2010). Data in Development: an Overview of Microdata on Developing Countries. IASSIST Quarterly. Winter/Spring 2010.

Ubaldi, B. (2013), Open Government Data : Towards Empirical Analysis of Open Government Data Initiatives. OECD Working Papers on Public Governance. No.22. OECD Publishing. *http://dx.doi.org/10.1787/5k46bj4f03s7*-en

West, Amy. (2010) Sources for International Trade, Prices, Production, and Consumption. IASSIST *Quarterly.Winter/Spring* 2010. Available Online: *http://www.iassistdata.org/downloads/iqvol334_341west.pdf*

Notes

1. Jungwon Yang, International Government Information and Public Policy Librarian. 240C Clark Library, Hatcher South, University of Michigan, Ann Arbor, MI 48109-1190. *yangjw@umich.edu*

2. http://www.census.gov/prod/www/decennial.html
3. http://www.stats.gov.cn/english/
4. http://www.stats.gov.cn/
5. www.cnki.nci
6. http://chinadataonline.org/
7. http://www.chinadatacenter.org/
8. http://chinadataonline.org/cge
9. http://citas.csde.washington.edu/data/data.html
10. SEDAC is one of the Distributed Active Archive Centers (DAACs) in the Earth Observing System Data and Information System (EOSDIS) of the U.S. National Aeronautics and Space Administration (NASA). SEDAC focuses on human interactions in the environment. Its mission is to develop and operate applications that support the integration of socioeconomic and Earth science data (http://sedac.ciesin.columbia.edu).
11. http://sedac.ciesin.columbia.edu/data/collection/cddc/sets/browse
12. http://chinadataonline.org/gambleapp/cityclient35/
13. Since 2006 the Statistics Korea has been integrating the KOSIS that are produced by statistical organizations according to the Project "Integrated National Statistics DB". Initially 462 kinds of national statistics from 113 organizations were being provided through KOSIS. In 2009 an additional 165 kinds of statistics from 44 organizations were integrated into the national statistics databases. The Project "Integrated National Statistics DB" was completed after integrating an additional 50 kinds of statistics produced by 6 organizations, including the Ministry of Public Administration and Security into the integrated database in 2010.
14. http://kosis.kr/eng/
15. For more detailed information about administrative division of South Korea, please use the following link of the Wikipedia (http://en.wikipedia.org/wiki/Administrative_divisions_of_South_Korea).
16. The list of available geosptial data files can be found in the land area image information service system website of NGII (http://air.ngii.go.kr/info/sub01.do).
17. https://www.open.go.kr/pa/PARetrieveInfoDisclosureGuide.laf?menuFlag=11
18. https://www.nsic.go.kr/ndsi/
19. http://gis.seoul.go.kr/SeoulGis/EnglishMap.html
20. http://gis.seoul.go.kr/SeoulGis/MetroInfo.jsp
21. http://gis.seoul.go.kr/Guide/Flex_map.jsp
22. http://imap.incheon.go.kr/icmap/map jsp?viewtheme=BASEMAP_AIREX
23. http://gris.gg.go.kr/
24. http://gis.gndo.kr/
25. http://map.gwd.go.kr/
26. http://www.stat.go.jp/english/data/nenkan/index.htm
27. http://www.stat.go.jp/english/data/chouki/index.htm
28. http://www.stat.go.jp/english/data/shihyou/index.htm
29. http://www.e-stat.go.jp/SG1/estat/ListEdo?bid=000001052147&cycode=0
30. http://www.e-stat.go.jp/SG1/chiiki/SelectMapDispatchAction.do#
31. http://e-stat.go.jp/SG2/eStatGIS/page/download.html#
32. http://e-stat.go.jp/SG2/eStatFlex/
33. http://www.stat.go.jp/english/info/guide/2011ver/05.htm
34. http://www.gsi.go.jp/kankyochiri/gm_japan_e.html
35. http://www.fas.harvard.edu/~chgis/japan/archive/
36. http://geodata.tufts.edu/openGeoPortalHome.jsp
37. The data sources information of China Geo-explorer is available when a user downloads a shapefile from MapExport section. The shapefile package contains a codebook which explains the data source.