由中美人口普查数据在开放利用方面的比较看中国大数据管理的缺陷

China's Defects in Big Data Management: A Comparison of China and United States Policy on Census Data Access

许晔 Dianna Xu

University of Wisconsin Madison

Toronto, Canada

March 17, 2017

由中美人口普查数据在开放利用方面的比较看中国大数据管理的缺陷 - 理念

2012年3月22日，美国奥巴马政府宣布投资2亿美元拉动大数据相关产业发展，将“大数据战略”上升为国家战略THE FEDERAL BIG DATA RESEARCH AND DEVELOPMENT STRATEGIC PLAN

奥巴马政府甚至将大数据定义为“未来的新石油”。

奥巴马政府用“新石油”比喻了大数据的重要性，但是没有涉及大数据不同于石油的特质。

从使用角度讲，目前人类使用石油的技术大多只停留在一时一地使用，很少能够重复使用。

大数据可以在同一时间在多地同时使用，并且可以无限期地重复使用。从这个意义上讲，大数据的使用价值极其可观。 -- 许晔

由中美人口普查数据在开放利用方面的比较看中国大数据管理的缺陷 - 历史

美国于1790年进行了第一次人口普查

是现代意义上第一个把人口普查写进宪法的国家

美国把人口普查数据归属于政府文件

美国法典第十七章第一节105款规定，政府文件（联邦政府）没有版权

中国的人口调查有近4000年的历史

中华人民共和国建立后，先后于1953、1964和1982年举行过3次人口普查。

根据《中华人民共和国统计法实施细则》和国务院的决定，自1990年开始改为定期进行，即每10年一次，在年号末位逢“0”年份举行。

由中美人口普查数据在开放利用方面的比较看中国大数据管理的缺陷 - 使用范围

在电子产品问世之前，美国政府通过遍布全国的近一万余家公立图书馆以及九百余家政府图书馆免费发送纸本人口统计资料。普通美国民众使用人口普查数据如同享用空气一样便捷。

可以说，美国政府为民众提供的免费的政府文件，包括人口普查数据是美国营造创新机制和土壤的了不起的举措。

20世纪初，福特先生的理念就是要让美国的每一个有工作的家庭都拥有一辆福特牌车。很难想像当年福特先生没有用到人口数据。

福特的汽车工业拉动了钢铁，铁矿，石油等等领域的一系列产业的发展。

近年来在美国领军的创新人物，大多都是在汽车房里起家的小人物。仔细观察他们把产品成功推向市场的带有科学依据的商业计划，无一例外，都会使用到人口普查数据。

由中美人口普查数据在开放利用方面的比较看中国大数据管理的缺陷 - 电子产品的使用范围

- 当下美国人口普查电子版数据种类繁多:

A. 人们可以从互联网上免费查询全美2000年和2010年的包括全国，洲，投票区，县，市，学区的各类人口数据 ，内容包括年龄，性别，就业状况，收入，宗教信仰，种族，受教育程度，残障，健康保险，等等。

B. 由740所大学的政治和社会研究联盟（ICPSR）提供的免费网络版全美人口普查数据库时间跨度回溯至1940年 。

C. 威斯康星大学麦迪逊分校（威大）购买了一款由橡树岭国家实验室出品的题为“全球人口土地扫描 ”（LandScan Global Population Databases）的数据库。这是一个世界范围的使用空间遥感和GIS图像分析，提供微小到一公里分辨率的全球人口分布数据库. $1,344美元/年

由中美人口普查数据在开放利用方面的比较看中国大数据管理的缺陷 – 版权和价格

美国LandScan Global Population Databases 这款数据库在版权页明确宣称“美国政府拥有一定版权”。又在服务对象及收费细则中规定联邦政府机构免费使用。

“社会浏览器”Social Explorer是一款具有强大检索功能的，囊括自1790年以来美国 220年的人口普查数据。包括25,000张地图， 400亿种类的数据元素和335,000变量。威斯康星大学支付“社会浏览器”的价格是$1,344美元/年

作为人口大国，中国历次人口普查都投入了巨大的人力和财力。不必说人口专家们为设计普查方案所贡献的智慧和付出的艰辛劳动，各级政府为这个庞大工程所做的数年准备，上千万普查工作人员的一系列培训活动直到登门入户，以及全国人民各个家庭为之作出贡献，统计结果最后再由专家们整理分析论证出版，人口普查数据是倾一国之力得到的大数据。

人口普查数据没有作为公共产品广泛开放利用，却成为个别部门牟利的垄断资源

中国人口普查数据的版权和价格

- **中国各地人口数据都表明拥有版权，部门利益丛生**
- 一种是以出版机构申明著作者，如本社，年鉴社等；第二类是以普查机构为名，如编委会，巴彦淖尔市第六次全国人口普查领导小组，呼和浩特市第六次全国人口普查领导小组；第三类的作者为 不详。
- 出版机构有国家统计出版社，省级，市级，地区的统计局或人口普查办公室
- 版权归属明确为出版社： 例如“安徽省2000年人口普查资料（上中下册）中国统计出版社出版发行，版权页宣称： 中国统计版图书，版权所有，侵权必究。

中国人口普查数据的版权和价格

笔者对中国2000年和2010年纸本人口统计出版物进行了对比析

- 从普查数据的销售情况看有几点值得注意：

1. 平均售价折合美金为$116美元/种。

2. 2010年比2000年的平均每种￥289元上涨￥449元，是2000年均价的2.6倍。

3. 根据威大2012年购书记录显示，尽管购入了几种大套书，每种中文书的平均价格为$19.77美元。相比之下，人口统计数据资料比普通书籍贵$96.23美元。

中国国家统计局人为制造的人口统计电子产品独家代理的垄断局面

- 中国2000年和2010人口普查数据电子版的海外销售昂贵，而且数据格式互相不能打通，人为背离大数据信息共享的理念

- 迄今为止，北美地区有两家公司出售中国人口数据电子版：

其一为“中国信息研究中心（China Data Center at the University of Michigan）”

收费原则依据卡内基高等教育委员会分类标准，以高校学生人数为依据进行分类（Carnegie Classification）。

其二为清华同方股份有限公司

中国大数据提升和发掘体系的制度缺陷

- 中华人民共和国版权法没有关于政府信息的条款
- 2007年国务院令第492号发布了中华人民共和国政府信息公开条例，条例的第二章“公开的范围” 第九条规定：“行政机关对符合下列基本要求之一的政府信息应当主动公开：（一）涉及公民、法人或者其他组织切身利益的；（二）需要社会公众广泛知晓或者参与的”。
- 人口普查是国家统一组织的，按国家法定的普查方案协调进行的专门性调查。这是一项带有强制性的国家工程，2010年人口普查还首次将我国境内的境外人员作为普查对象。如此重要的政府信息自然需要社会公众广泛知晓或者参与使用，理当列入主动公开的信息。
- “政府信息应当主动公开”的规定只是柔性倡导，并没有与版权法结合，因而没有刚性的法律制度保障。

中国大数据提升和发掘体系的制度缺陷

- 由此带来的后果就是民众作为受众在与政府之间的关系中处于被动地位。政府作为全国最大的信息生产，收集，使用和发布单位仍然可以用国家赋予的权力，把用纳税人的钱收集到的信息变成商品投放市场，从公民手中谋利。
- 中国的惨痛教训是2010年人口普查资料的出版发行，反映了国家2007年颁布的政府信息公开条例并未得到各级出版机构的贯彻和执行。具有行业性的行政资源成为其下属出版机构牟利的垄断资源。
- 2007年国务院令第492号发布的中华人民共和国政府信息公开条例只有柔性倡议，没有刚性的法律约束。

中国大数据管理缺乏反垄断意识：政府机构参与独家授权，监管部门没有监管

- 国家统计局将2000年和2010年的 全国人口普查数据电子版的开发销售权独家出售给上述公司的行为，恰恰吻合《中华人民共和国反垄断法》（2007年8月30日通过）第一章总则 第三条 “本法规定的垄断行为包括：（一）经营者达成垄断协议；（二）经营者滥用市场支配地位；……” 然而目前独家授权仍然普遍发生，不但反映了出版发行领域无视反垄断法的行为，也反映了一种集体的，带有普遍性的对反垄断要义认识的缺失。
- 国家统计局对2000年和2010年的全国人口普查数据的出版、发行以及海外推广完全委托部门所属的国家统计出版社，竟然完全没有意识到这是一种违法行为。

核心问题 – 中国人口数据因为**行政垄断**而没有得到有效的公开利用是中国数字出版问题的典型缩影

中国人口普查数据这款具有巨大社会文化价值的信息产品的开发和利用迄今处于极其低级的一个水平。这是中国数字出版的一个典型缩影：受制于各种制约，导致内容资源稀缺，基本不具备市场竞争能力。

自2010年至今，六年时间过去了，对于耗费巨大人力、物力资源，通过层层行政动员获得的中国人口普查数据在出版、发行等后续社会服务效果的评估方面，国家相关监管部门竟然无人过问。

这表明无论在执政理念上，还是在法律制度建设方面，中国的国家治理能力离现代化的目标还有很大差距。

核心问题 – 中国人口数据因为行政垄断而没有得到有效的公开利用是中国数字出版问题的典型缩影

- 由此带来的后果就是民众作为受众在与政府之间的关系中处于被动地位。政府作为全国最大的信息生产，收集，使用和发布单位仍然可以用国家赋予的权力，把用纳税人的钱收集到的信息变成商品投放市场，从公民手中谋利。

- 核心问题： 反映了集体性的对民众在大数据应用中的权利，地位和作用缺乏应有的认识。

由中美人口普查数据在开放利用方面的比较看中国大数据管理的缺陷 – 电子产品

- 威斯康星大学支付同方知网$4,746 购买中国2010年人口普查数据库（省级包括少数几个市的不带地图）的数据库。
- 垄断价格毫无疑问需由消费者承担。这款数据库原价$7,910威斯康星大学参加了集团购买，按40%减价支付。这款产品仅包括省级（还有少数市级）的数据。虽然题为“第六次全国人口普查的数据库”实际上包括了一些2000年及2000年之前的普查数据。比如，全国范围13种当中，有4种是2010年的数据，其余都是2000年及2000年之前的数据。环渤海地区的9种当中，有 5种是2010年的数据，其余都是1990年的数据。这款数据库总共包括134种数据，是以电子书的方式销售，并且不带地图。

同方和中国信息研究中心两家公司的产品采用不同的理念和处理方式组织数据

- 即使用户舍得花血本买下两套系统，也必须自己打理断裂式的数据表述系统

- 国家统计局的每隔十年一次的短期独家授权行为完全没有从大数据的提升和挖掘着眼，从技术层面分析也难以理解。

- 中国在世界性的大数据的提升和发掘战略背景下，其核心障碍是行业条块分割，部门利益丛生，表明无论在执政理念上，还是在法律制度建设方面，国家治理能力离现代化的目标还有很大差距

由中美人口普查数据在开放利用方面的比较看中国大数据管理的缺陷 - 结束语

1. 中国在数据领域的管理问题反映出长期积累的思维定势，观念陈旧，体制障碍，对公众的投入没有从法律上给予版权保护。

2. 大数据时代呼唤着一个更加开放的，民主的和公平的社会管理制度。开放存取 ，开放数据，开放教育正在全球形成新的浪潮。

3 . 大数据管理的一大特点就是用统一的数据标准，打破信息孤岛，实现跨行业跨部门的大数据集成，提供整体服务系统等等。

4. 中国急需出台纲领性的立法原则，厘清公民的权利和政府的义务，设计出一套行之有效的地方和各级政府通力配合的管理大数据的体制

由中美人口普查数据在开放利用方面的比较看中国大数据管理的缺陷 - 结束语

5. 在人口普查数据普及利用方面，中国已经比美国落后了二百二十六年。

6. 当下中国政府和人民都在努力提升中国的创新机制，打造智慧城市智慧家园。

7. 然而，我们守着“人口普查大数据”这座应有尽有的金山银矿却任由少数几家出版机构垄断数据，其损失无可计量。

8. 中国的出版界，法律界和各界民众必须通力合作，以只争朝夕的精神解决大数据开发利用的核心问题。

Thank you!　谢谢！

Dianna Xu

References

U.S. Code (2012 edition) http://www.gpo.gov/help/index.html#about_united_states_code.htm)Title 44, Public Printing and Documents Chapter 19, Depository Library Program Section 1901

References

- § 1901. Definition of Government publication“Government publication” as used in this chapter, means informational matter which is published as an individual document at Government expense, or as required by law. (Pub. L. 90–620, Oct. 22, 1968, 82 Stat. 1283.)

http://www.gpo.gov/fdsys/pkg/USCODE-2012-title44/pdf/USCODE-2012-title44-chap19-sec1901.pdf

References

- § 105 . Subject matter of copyright: United States Government works
- Copyright protection under this title is not available for any work of the United States Government, but the United States Government is not precluded from receiving and holding copyrights transferred to it by assignment, bequest, or otherwise.
- It is not subject to copyright in the United States and there are no copyright restrictions on reproduction, derivative works, distribution, performance, or display of the work.

- Census 2000 Gateway - U.S. Census Bureau - Census.gov
- http://www.census.gov/main/www/cen2000.html

References

Census of Population, 1940 [United States]: Public Use Microdata Sample (ICPSR 8236)

http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/8236

http://web.ornl.gov/sci/landscan/

http://web.ornl.gov/sci/landscan/landscan_data_avail.shtml

Who can have access and how is access obtained?

LandScan™ Data are available free of charge for U.S. Federal Government agencies.

http://www.socialexplorer.com/