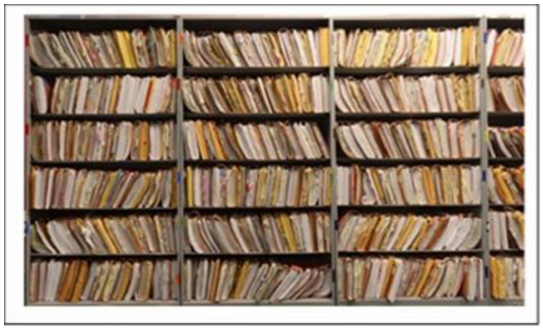
识别码在大数据时代的要义

胡善庆博士 [1], 丁浩 [2]

[1] 美国联邦政府退休官员，乔治•华盛顿大学教授，Jeremy.s.wu@gmail.com

[2] 华盛顿大学访问学者， edwarddh101@gmail.com

First published as a personal blog on April 30, 2013.



摘要

在 21 世纪，大数据承诺将为社会有效治理以及大众信息分享做出贡献。尽管任何数据本身都包含一定的信息与作用，但是关联和整合后的数据不仅减少收集数据的重复性，而且极度的增加它的價值和可用性。识别码在这个过程中不仅促进实际记录和数据的整合，而且是解放大数据威力的关键。如果识别码没有得到正确的使用和管理，它亦将会是系统失灵、误用和滥用、甚至欺诈及犯罪的元凶。因此，除了技术以外，合理的统计学设计，提高质量的反馈，适当的教育和培训，相关的法律法规，公众的认知，这些都将成为识别码和大数据有效和负责任应用的必要条件。

识别码的必要性

在学生入学时，会有档案存储学生的各种数据，比如：姓名、性别、年龄、家庭背景、专业等。当学生选修一门课并获得成绩时，这个结果也被记录下来。当这个学生满足了所有毕业要求，另一条记录会显示出她的加权平均分并且获得的学位。

每一条记录都是这学生的一个“快照”，随时间累积成为行政记录。这些纵向“快照”提供每个学生受教育情况的丰富信息。

当学生进入工作单位，更多的关于她工作的数据将被收集，伴随她一生，这些数据包括：她的职业及工作单位，工作表现，工资及晋升情况，保险和税的支付数额，就、失业状态等。

在同样的情形下，大量关于公司的数据也会被收集。这些数据记录了：最初注册成立，收支财政状况报告，上市情况，收购或者与其他公司合并，所缴税费，收入增长和雇员增加，公司的扩增或是公司的倒闭。

这些行政记录过去被封存于满身尘埃的文件柜里，但是在千禧年伴随着大数据时代的到来，它们大部份都已数字化。

对学生数据及时和适当的整合将会提供空前的细节，使我们更详细地了解这所学校运作情况，比如说毕业率随时间的变化。当数据整合扩展到所有学校，我们将更好的了解这个国家的教育状况，例如它对就业和经济增长的潜力和支持。这些就是 21 世纪大数据承诺将会给我们带来的变化。从分配资源，评估表现，到制定政策，社会的方方面面都可以从大数据的细节和深度中促进社会有效治理及大众信息分享。

尽管任何数据本身都包含一定的信息与作用，但是关联和整合后的数据将更为重要，因为它不仅减少收集重复的数据，而且极度的增加它的價值和可用性。识别码的重要意义是促进实际记录和数据的整合。在这个过程中，统计学家可以作出卓越的贡献，运用他们的智慧和知识创立新的统计系统。

识别码的种类

当文件例如纸质表格还未被数字化前，人名或者公司名称是被常用的识别码。通常来说，人们会用相同的名称整合记录并给他们排序，比如英文的字母，中文的笔画，或按时间顺序。

但是，使用名字的一大弊端是他们并不是独特唯一，特别是在电脑大量处理数据时，这一弊端尤其明显。据 2006 年的统计，李、王、张、刘这四个中国最大姓氏占了 3.34 亿人口[1]，超过美国人口总数。同样的中文姓名，也有繁体和简体中文的可能分别。英文名罗伯特（Robert）有至少七种不同使用方法，包括：Bert，Bo，Bob，Bobby， Rob，Robbie，以及 Robby，它在 2011 年美国出生的男性人名中的使用率排第 61 位[2,3]，而 Bert 又可以是英文名 Albert 的缩写。个人又有可能更改名字或者有不止一个名字；女性可能在结婚后改名。人为的错误又可能增加不正确的名字。在使用不同语言的情况下，引用同一名字更是特别困难。

在注册的过程中，公司的名称会被检查以确保不出现重名。公司的名称包括它的商标也会被当地，全国性以及国际性的规则和法律受到保护。但公司仍有可能使用多个名称，包括它的缩写和公司股票代码，而且它也有可能在合并，重组，被收购的时候变更名称，或者只是简单的更改品牌。

非唯一的识别码会造成不正确链接和合并数据的风险，导致不正确的结果或结论。虽然给一个名称增加辅助信息，比如说年龄，性别和地址，可以减少风险，但是并不能完全去除错误配对记录和数据的可能性，而且会增加处理数据的时间。

识别码可以由一系列的数字、字母或特殊字符(字母数字)组成。越来越多使用单纯的数字来组成识别码，应用于现代的机器排序，链接和合并电子记录。因为纯数字识别码不依赖于语言系统，受到比较少的限制。使用字母数字的识别码，可能适合使用拉丁语系的系统，但是那些非拉丁语系的系统就比较难以使用、明白或理解。同时，数字字符比较字母数字字符容易排序。

当美国在 1935 年通过社会安全法案时，履行法案遇到的第一个挑战就是创造如何“永久识别每个个体”的识别码，同时保有“以后能够有效和无限制的增加识别对应增长工人的功能”[4]。一个八位字母数字系统最

先被提出来，但很快遭到了统计机构、劳工及法律部门的反对。这个变换被描述为“机器会如何深远地映响[政府] 操作”的第一个徵兆[4,5] 。这些都是在计算机实际使用以前发生的事情。

如今，信息科技的巨大影响很明显，不但是政府，商业和个人活动的方方面面，而且影响力还在不断增强。一个识别码可以应用于个人，一家公司，一辆车，一张信用卡，一箱货物，一个电子邮箱账户，一个地方，或者是任何一个实际个体。

如果一条电子记录不包含识别码，或不能和其他记录連接，大数据中称为缺乏“结构”或者叫做“无结构”。从 21 世纪初开始，“无结构”数据比有“结构”数据出现的频率多得多，但是它们比有“结构”数据包含较少信息内容，也更难应用，特别是在社会和经济方面，我们很难得到后续、连贯和可靠的时序信息。

如何有效地使用识别码将是发挥大数据巨大作用的关键。

有效使用识别码

1. 匹配和合并记录。理想的识别码同时互斥，完全穷尽，在代码和实体间建立了明确的一一对应关系，同时也会延生到未来的记录。识别码促进对电子记录直接有效地排序、匹配及合并，具有无限扩增实体信息内容的潜能。

2. 匿名和保护身份。因为代码是实体的匿名，所以它为身份保护提供第一道防线。但随着识别码重要性的增强，以及它与其他数据链接相对容易，通过识别码伪造及盗用身份的危险性和可能性也在增加，这就需要加强对识别码的政策和负责任的管理，以使它起到保护的作用。

3. 基本描述和分类。识别码可以对数据内容提供最基本的描述，迅速从中得到简单的信息或者是总结。随着时间的推移，这个概念延伸到识别码的分类和“元数据”的发展[6,7]，这个过程包含了在数据系统中建立有效的结构以及扩展它们跨系统的应用。

4. 初部质量检查。无意的人为输入错误以及不正确的转录识别码都可能对整体数据和最终分析结果的质量造成破坏。欺诈或恶意改变识别码亦可能会造成对数据的完整性和可靠性严重的破坏。“效验码”[8,9]在早期检查中的使用可使识别码中常见错误降低 90%。

5. 促进统计学创新。通过对每个学生数据连续不断的收集和整合，可以建立一个含有所有学生和学校丰富信息的动态框架。在严格保护个体隐私和数据安全的同时，新的变数可以定义用做分析研究，描述一所学校的表现或者是一个国家教育状况的统计结论可以是定时或实时产生。在美国及中国，建立这些动态框架和纵向数据系统都已起步[10]。Data Quality Campaign[11]列“唯一州际连接学生数据和主要数据库学生识别码”为建立全美教育纵向数据系统中最关键的部分。

美国和中国的个人识别码

美国没有全国性的个人识别系统。社会安全号创立于 1936 年，还在商业使用电脑之前，用于追踪劳工的收入。在电脑大规模使用以后，社会安全号作为识别码表现出一些优势和劣势，如图 1。

图 1 美国识别码

识别码是九位的社会安全号由三部分组成：

地区号码（三位）- 最初是发放社会安全号的地区代码，后来代表申请邮寄地址的邮政代码

组号（两位）- 代表着一个社会安全号集合被指定为一个组

系列号码（四位）- 从 0001 到 9999

社会安全号的申请过程中[12]收集人口信息，包括名字、出生地、出生日期、国籍、种族、性别、父母的名字和社会安全号、电话号码和邮政地址。美国社会安全部负责社会安全号的发放。有一些社会安全号被保留，没有使用。一旦一个社会安全号被发放，它应是唯一的，因为它不会被第二次发放。但重复的情况仍可能存在。

1938 年，一个钱包的生产厂商显示他们的产品是多么适合社会安全号卡来促销其在百货商场出售的钱包，但是他们使用一张自己员工的社会安全号卡[13]。这导致有四万人错误的使用了这个社会安全号，甚至到 1977 年还有人将这个号码做作为自己的社会安全号。

自从社会安全号的产生，它被政府部门和私有企业的使用显著增加。从 1943 年开始，总统行政命令要求各联邦政府部门必须使用社会安全号建立拥有永久账户号码的系统[5]。在 1960 年代初，政府雇员和个体报税者必须使用社会安全号。到 1960 年代末，社会安全号被作为军人的识别码。在整个七十年代，当电脑被越来越多使用后，金融活动，如开设新银行账户和申请信用卡和贷款，以及联邦福利的运行中，社会安全号成为必不可少的一部分。从 1986 年开始，如果父母想要有受抚养人的免税，就必须将其抚养人的社会安全号也列在税表里。在法律实施的第一年，这反欺诈行动减少了七百万的受抚养人数[14]。

社会安全号可以将同一个人的很多电子文件链接合并到一起，因此它本质上作为了非官方全国性识别码，但是它也可能直接造成误用或者滥用，例如身份盗用[15]。社会安全号没有效验码，它并不能有效的作为身份的認证。有学者也展示如何用公开的信息“异常精确的重建社会安全号”[16]。这些年

图 2 中国相对较晚开始使用个人识别码

在美国，识别码的这些脆弱点使得人们更加小心谨慎和负责任的使用社会安全号。1943 年要求使用社会安全号的行政命令也被废除，取而代之的是在 2008 年颁布的行政命令使社会安全号成为可以选择而非必须的。

中国相对较晚开始使用个人识别码，如图 2。在 1999 年 7 月 1 日，身份证号码由 15 位提升为 18 位，其中出生年份由两位变为四位，并且增加效验码。18 位身份证号由四部分组成[17,18]：

识别码的要义地区代码（六位）—个人住址的行政编号

生日代码（八位）—按生日的年月日顺序组成

系列代码（三位）—其中奇数代表男性，偶数代表女性

效验码（一位）—使用 ISO 7064 标志算法，基于前面 17 位数字计算得到[18,19]

居民身份证由居民常住户口所在地的县级人民政府公安机关基于未满 16 岁居民的申请签发。居民身份证登记的项目包括：姓名，性别，民族，生日以及居住地址。居民身份证有效期长至永久，也可能短至五年，取决于申请人的年龄。根据官方的声明，居民身份证号在中国电子健康档案中也用于记录个人的健康信息[20]。

中国及美国的商业识别码和工业分类码

美国企业的僱主识别码相当于个人的社会安全号[21]。但是，这里的企业包括地方，州和联邦政府，也包括无雇员的公司，亦包括需要为其雇员缴纳税款的个人公司。僱主识别码是由美国税务局负责指派的一个九位数字，它的形式是 GG-NNNNNNN，其中 GG 在 2001 年前是公司所在地的代码，而后七位数字没有特别的含义。一旦一个僱主识别码被使用，美国税务局就不会再次使用。另外，每个州亦各有自己的僱主识别码用于税务收缴和行政管理。

在联邦僱主识别码的申请过程中有以下信息被收集:正式名称，交易名称，法人姓名，责任人员，邮政地址，商业地址，公司类型，申请原因，成立时间，财政年度，未来 12 个月员工数目估计，首次工资发放日期以及公司主营业务[22]。

美国统计部门使用北美工业分类系统（以下称 NAICS）来对公司营业进行归类，以期能收集，分析以及发布美国经济的统计信息[23]。在 1997 年，北美工业分类系统（NAICS）继承取代工业标准分类系统(SIC)。

NAICS 是一个层级分类代码系统，其中可能包含有 2 到 6 位数字。最高层级的 2 位数字代表主要经济部门，例如建筑和生产。每个 2 位数字所代表的部门都包含一系列的 3 位数字子部门，而它又包含有一系列 4 位数字的工业集团。例如 31 到 33 是表示生产部门，而碾米工业在其所属层级之中：

311 食品加工制造业
3112 粮食和油菜籽加工业
31121 面粉和麦芽生产业
311212 碾米业

层级系统其中的一个优势就是它可以相当容易地链式聚集产业总值。比如说，所有代码为 311x 企业的总和就组成了代码为 311 的食品加工制造业。

持续用 NAICS 代码準确地把企业分类是一大挑战，因为在当今快速变化的动态国际经济环境下，一夜之间过时的行业会被淘汰，

新的行业也会出现成长，过去的“高科技企业”及最近的“绿色”行业就是例子之一。使用 NAICS 代码的过程中有着理解和持续性的问题，例如美国统计局和美国劳工统计局就因为数据来源和 NAICS 代码分类不同，令到各自创立和维护的商业框架有異[10]。不一致使用 NAICS 代码破坏甚至造成对时间序列和纵向数据分析无效。

中国的新企业必须向当地质量监督局申请 9 位数的国家组织机构代码，其中由 8 位数字（或大写拉丁字母）本体代码和 1 位数字（或大写拉丁字母）效验码组成。中国的组织机构代码，借鉴原 ISO6523《数据交换标识法的结构》（现 ISO6523《信息技术组织和组织各部分标识用的结构》）国际标准的基础上，根据 GB 11714—1997《全国组织机构代码编制规则》国家标准的规定，编制的全国统一的组织机构代码识别标识码[25]。可以通过网上信息核查系统基于国家组织机构代码查询组织机构的信息[26]。

国内及国外的经济学家和其他学者十分认可中国工业企业数据库的价值。透过相当的投资，这个丰富的综合数据系统从 1998 年开始纵向描述中国差不多所有的国有和大型企业（2010 年前销售额在 500 万人民币以上及 2010 年后销售额在 2000 万人民币以上的企业）。但是，十分严重的质量问题已有所报导，而主要数据错误原因可以追溯到不正确和不连贯地使用识别码[27]。虽然中国从 1989 年就开始标准化国家组织机构代码，并且现在已经进行到了第三阶段，但是这个问题仍然存在[28]。

就在上个月，广东省宣布他们运用国家组织机构代码这个平台推动反腐败[29]。中国也有一个根据 GS-T4754-2002 文件而建立的标准工业分类系统[30]。这个层级系统有四类，其中最高层为一个字母，其余分别有 2 位、3 位、4 位数字代码表示较低层级。以前述的的碾米业为例，中国的分类系统中表示为以下层级：

C　　制造业
C13　　农副食品加工业
C131　　谷物磨制
C1312　　大米加工企业

总结

随着科技的轉变和发展，收集大规模的数字化数据的成本将更低，速度也将更快。这些是大数据时代的标志。

这些大数据包含了空前规模的信息。如果数据整合和结构化，它们的价值和功能将会暴涨，超过现有数据系统所能提供。识别码促进数据的链接和合并，是提供这些巨大机会的关键。

识别码能解放大数据的巨大能量。如果我们不能正确使用和管理识别码，它同样可以成为系统失灵，误用和滥用，甚至是欺骗和犯罪行为的罪魁祸首。

现实使用识别码的挑战是多样复杂。除了科技技术，统计学设计和质量回馈途径，适当的教育和培训，有效的政策和监控，以及公众的意识參与都是有效负责使用识别码所必须的。未来的文章中将讨论这些话题。

参考文献

[1] 360doc.com. Quantitative Ranking of Chinese Family Names (中國姓氏人口數量),November 25, 2012. Available at

http://www.360doc.com/content/12/1125/17/6264479_250155720.shtml on April 29, 2013.

[2] Wikipedia. Robert. Available at http://en.wikipedia.org/wiki/Robert on April 29, 2013.

[3] U.S. Social Security Administration. Change in Name Popularity. Available at http://www.ssa.gov/OACT/babynames/rankchange.html on April 29, 2013.

[4] U.S. Social Security Administration. Fifty Years of Operations in the Social Security Administration, by Michael A. Cronin, June 1985. Social Security Bulletin, Volume 48, Number 6. Available at http://www.ssa.gov/history///cronin.html on April 29, 2013.

[5] U.S. Social Security Administration. The Story of the Social Security Number, by Carolyn Puckett, 2009. Social Security Bulletin, Volume 69, Number 2. Available at http://www.ssa.gov/policy/docs/ssb/v69n2/v69n2p55.html on April 29, 2013.

[6] Wikipedia. Metadata. Available at http://en.wikipedia.org/wiki/Metadata on April 29, 2013.

[7] Wikipedia. 元数据. Available at http://zh.wikipedia.org/wiki/元数据 on April 29, 2013.

[8] Wikipedia. Check Digit. Available at http://en.wikipedia.org/wiki/Check_digit on April 29, 2013.

[9] Wikipedia. 效验码. Available at http://zh.wikipedia.org/wiki/校验码 on April 29, 2013.

[10] Wu, Jeremy S. 21st Century Statistical Systems, August 1, 2012. Available at http://jeremyswu.blogspot.com/2012/08/abstract-combination-of-traditional.html on April 29, 2013.

[11] Data Quality Campaign. 10 Essential Elements of a State Longitudinal Data System.Available at http://www.dataqualitycampaign.org/build/elements/1 on April 29, 2013.

[12] U.S. Social Security Administration. Application for a Social Security Card, Form SS-5.Available at http://www.ssa.gov/online/ss-5.pdf on April 29, 2013.

[13] U.S. Social Security Administration. Social Security Cards Issued by Woolworth. Available at http://www.socialsecurity.gov/history/ssn/misused.html on April 29, 2013.

[14] Wikipedia. Social Security Number. Available at http://en.wikipedia.org/wiki/Social_Security_number on April 29, 2013.

[15] President's Identity Theft Task Force. 2007. Combating Identity Theft: A Strategic Plan. Available at http://www.idtheft.gov/reports/StrategicPlan.pdf on April 29, 2013.

[16] Timmer, John. New Algorithm Guesses SSNs Using Data and Place of Birth, July 6, 2009. Available at http://arstechnica.com/science/2009/07/social-insecurity-numbers-open-to-hacking/ on April 29, 2013.

[17] baidu.com. GB11643-1999 Citizen Identity Number 公民身份号码. Available at http://wenku.baidu.com/view/4f19376348d7c1c708a14587.html on April 29, 2013.

[18] Wikipedia. Resident Identity Card. Available at http://en.wikipedia.org/wiki/Resident_Identity_Card_(PRC) on April 29, 2013.

[19] Wikipedia. ISO 7064. Available at http://en.wikipedia.org/wiki/ISO_7064:1983 on April 29, 2013.

[20] baidu.com. Electronic Health Record 电子健康档案. Available at http://wenku.baidu.com/view/348d5a18a300a6c30c229fec.html on April 29, 2013.

[21] Wikipedia. Employer Identification Number.Available at http://en.wikipedia.org/wiki/Employer_identification_number on April 29, 2013.

[22] U.S. Internal Revenue Service. Form SS-4: Application for Employer Identification Number. Available at http://www.irs.gov/pub/irs-pdf/fss4.pdf on April 29, 2013.

[23] U.S. Census Bureau. North American Industry Classification System.Available at http://www.census.gov/eos/www/naics/index.html on April 29, 2013.

[24] National Administration for Code Allocation to Organizations. Introduction to Organizational Codes, 组织机构代码简介. Available at http://www.nacao.org.cn/publish/main/65/index.html on April 29, 2013.

[25] Wikipedia. ISO/IEC 6523. Available at http://en.wikipedia.org/wiki/ISO_6523 on April 29, 2013.

[26] National Administration for Code Allocation to Organizations. National Organization Code Information Retrieval System, 全国组织机构信息核查系统. Available at http://www.nacao.org.cn/ on April 29, 2013.

[27] Nie, Huihua; Jiang, Ting; and Yang, Rudai. A Review and Reflection on the Use and Abuse of Chinese Industrial Enterprises Database. World Economics, Volume 5, 2012. Available at http://www.niehuihua.com/UploadFile/ea_201251019517.pdf on April 29, 2013.

[28] National Administration for Code Allocation to Organizations. Historical Development of National Organization Codes, 全国组织机构代码发展历程. Available at http://www.nacao.org.cn/publish/main/236/index.html on April 29, 2013.

[29] National Administration for Code Allocation to Organizations. Guangdong Aggressively Promotes the Use of identification Codes in its Campaign against Corruption, 广东积极发挥代码在反腐倡廉中的促进作用, March 7, 2013. Available at http://www.nacao.org.cn/publish/main/13/2013/20130307150216299954995/20130307150216299954995_.html on April 29, 2013.

[30] baidu.com. National Economic Industry Classification, GB-t4754-2002, 国民经济行业分类(GB-T4754-2002)(总表). Available at http://wenku.baidu.com/view/69f04af8c8d376eeaeaa31cf.html on April 29, 2013.

作者简介：

胡善庆博士,美国联邦政府退休官员。现任乔治•华盛顿大学统计系客座教授及上海华东师范大学大数据创新中心主任。邮件：jswu@gwu.edu

丁浩，现为华盛顿大学访问学者及数据治理促进会成员，毕业于华中科技大学和乔治华盛顿大学。