数据自治开放模式

朱扬勇[1,2]，熊赟[1,2]，廖志成[1,2]，叶雅珍[1,2,3]
1. 复旦大学计算机科学技术学院，上海 201203；2. 上海市数据科学重点实验室，上海 201203；
3. 东华大学计算机科学与技术学院，上海 201620

摘要
介绍了现有数据资源管理模式，并指出了其中存在的问题；提出了数据自治开放模式，并介绍了数据自治开放技术。数据自治开放模式有望成为数据开放的基本模式，是政府数据开放共享、企业及个人数据交易、国家数据主权实现的一种可行方法。

关键词
数据开放；数据共享；数据自治开放；大数据

中图分类号：TP311　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2018013

Self-governing openness of data

ZHU Yangyong[1,2], XIONG Yun[1,2], LIAO Zhicheng[1,2], YE Yazhen[1,2,3]
1. School of Computer Science, Fudan University, Shanghai 201203, China
2. Shanghai Key Laboratory of Data Science, Shanghai 201203, China
3. School of Computer Science, Donghua University, Shanghai 201620, China

Abstract
Existing data management techniques and their drawbacks were introduced. A novel model called self-governing openness of data was presented, and the details in using the technique were also illustrated. Self-governing openness of data would be a basic path for realizing the openness of data. Meanwhile, it should be an applicable approach for the openness and sharing of government data, the transaction of enterprise and personal data, and the realization of national data sovereignty.

Key words
open data, data sharing, self-governing openness of data, big data

1 引言

数据已经成为国家基础性战略资源，推动数据资源开放共享是国家《促进大数据发展行动纲要》的核心内容。政府和公共数据资源的开放共享不仅是构建一个透明的政府，更重要的是创造新兴战略产业（数据产业），推进传统产业转型升级，成为驱动创新的主要因素[1]。但在实施过程中，数据开放共享面临着“数据拥有者不愿、不敢、不会开放共享”的问题。由于数据可以以极低成本复制和传播（复制一份数据的成本远远低于生产一份数据的成本），加之数据资源的战略性和商业价值越来越显现，这就导致生产数据的意愿远远低于复制数据的意愿，因而呼吁数据开放的人越来越多；数据生产者越来越不愿意将其拥有的数据开放出来。因此，数据资源的开放变得越来越难以实现。事物的价值源于稀缺性，由于数据可以几乎零成本地复制和传播，所以，数据开放意味着数据资源的稀缺性丧失，从而丧失其原有的价值，这是不愿意开放数据的根本原因。因此，如何既开放数据又保持数据资源的稀缺性，是一个亟待解决的重大问题。本文围绕该重大问题，探索数据开放模式，提出数据自治开放模式。

开放数据是指数据免费开放给每一个希望使用数据的人，主要是指政府和公共数据资源应该开放给公众，使公共数据能被任何人、在任何时间和任何地点自由利用、再利用和分发[2]；数据共享是指对数据使用对象、使用时间和使用地点加以限制，主要是对使用对象进行限制，即将数据开放给特定对象[3]，只有特定对象在特定的时间、地点使用指定的数据，可以理解为开放数据的限制版；数据交易是指数据拥有者依据法律在市场交易规则下进行自由交易。总体而言，开放数据、数据共享和数据交易都是数据拥有者将数据开放给数据使用者，只是在范围、对象、是否收费等方面有所不同，三者面临的核心问题都是“数据如何治理”，具体说就是“如何控制数据使用者传播或滥用数据”。开放数据模式不对开放出去的数据进行治理[4]；数据共享模式则由共享圈共同治理数据，但共享圈约束有限，数据常常流出共享圈而造成事实上的开放数据；数据交易的数据治理目前还没有具体做法。因此，为了方便叙述，将开放数据、数据共享和数据交易统称为“数据开放”。

当前，绝大部分数据资源都还处在封闭不开放的状态，数据完全由数据拥有者治理。拥有者尽量保护系统数据不受外界侵害，不对外界开放，即数据是自治封闭的。现有的数据资源管理技术（数据库管理系统、文件系统）和应用软件技术都支持这种数据自治封闭模式。数据自治封闭模式的问题是数据资源只能由数据拥有者使用，没有发挥数据资源应有的价值。要更大程度地开发利用数据，就需要将数据资源开放出来。然而，数据会被怎样开发利用事先可能是不知道的，使用数据的软件也是事先不知道的、基本外部的、数量无限的、安全不可控的、隐私不可控的。现有的数据库管理系统软件根本无法处理数据开放的应用需求，因此，需要探索新型的数据资源管理技术和数据开放模式。

本文提出的“数据自治开放”是指数据拥有者在法律框架下对数据进行自行确权和管理、自行制定开放规则（即数据自治），然后将数据开放给使用者，包括上传到数据应用软件使用数据和下载数据到使用者的设备中（使用者没有数据

治理权）。

数据自治开放模式有望成为数据开放的基本模式，是政府数据开放共享、企业及个人数据交易、国家数据主权实现的一种可行方法。

2 数据资源开放与稀缺性的矛盾

数据资源是重要的现代战略资源，而且其重要性越来越显现，在21世纪有可能超过石油、煤炭、矿产，成为最重要的人类资源。提高数据资源开发利用水平、保护国家的战略资源是增强我国综合国力和国际竞争力的必然选择[5]。2011年5月麦肯锡公司发布的《大数据：下一个创新、竞争和生产力的前沿》报告、2012年2月世界经济论坛年会发布的《大数据，大影响》报告等，都突显了大数据的价值和重要性①②。2012年3月美国白宫科学和技术政策办公室发布《大数据研究和发展倡议》，标志着美国率先将大数据上升为国家战略③，随后，日本、法国、澳大利亚、英国等国家也开始发布大数据国家战略或计划等。2014年和2015年，我国将大数据写入《政府工作报告》，2015年8月19日国务院通过《关于促进大数据发展的行动纲要》，2015年10月26—29日中国共产党第十八届五中全会提出“实施国家大数据战略”等，这些都表明数据已经成为重要的战略资源。

一份数据资源的价值除了体现在它的内容外，更重要的方面体现在它的稀缺性。内容再重要的数据资源，如果人手一份或者随时可以获得，那就没有人愿意付费购买，其本身的价值就难以体现出来。由于数据可以以极低的成本复制和传播，所以一旦数据资源生产者将数据资源开放，就意味着该份数据资源可能会传遍世界，从而丧失稀缺性。

矛盾在于：如果不开放，则数据资源只能自用，价值发挥有限；如果开放，则数据资源可能丧失稀缺性，使数据资源生产者丧失利益。

由于目前在技术上还没有保持数据资源稀缺性的数据开放技术，所以，在实践中，数据资源拥有方不愿开放、不会开放就成了数据开放进程中的“拦路虎”。另外，由于政策制约，数据资源拥有方还存在不敢开放的问题。

不愿开放：指数据资源拥有者不愿意在没有获得足够利益的情况下进行数据开放。

不会开放：指尽管数据持有者希望将数据资源开放出来，但是由于现行技术并不适合数据资源开放，所以不知道如何实现数据资源开放。

不敢开放：指怕承担责任，目前的政策是“谁有数据谁负责”，因此，万一数据开放出了问题，数据拥有部门就要承担责任；另外，一些数据拥有者担心数据开放后，数据资源的稀缺性会丧失。

3 现有数据资源管理模式

现行的数据管理技术是面向数据自治封闭的，不适合数据开放共享，急需开发面向数据开放共享的技术。

3.1 政府开放数据

政府开放数据的典型代表是2009年美国政府推出的网站www.data.gov，因此，2009年一般被认为是数据开放元年。之前是政府信息公开，政府向公众公开各种报告、决策结果；政府开放数据是信息公开的进一步，即将形成报告和决策的原

① https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/ MGI_big_data_full_report.ashx

② http://www.weforum.org/reports/big-data-big-impact-new-possibilities-international-development

③ https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf）

始数据也公开，主要内容是政府应该向公众透明[6]。2015年我国国务院印发的《促进大数据发展行动纲要》明确提出，数据开放共享主要是指政府和公共数据资源应该开放给公众共享。

从国际上看，政府数据开放主要通过制定战略或政策文件形式指导开放，又因涉及多个部门，往往由最高领导层发布，例如美国前总统奥巴马在2009年和2013年两次发布开放政府数据的行政令；英国在2010年和2011年先后两次发布《致政府部门开放数据函》等。开放过程中，各国通常把数据作为一种国家资产进行管理，要求建立相关的制度。比如，建立数据资产目录，各部门需梳理数据资产，明确各类数据的开放属性（公开、限制公开、不公开）；建立数据开放的目录，确定哪些是已开放的，哪些是将来会开放的。并且，目录保持持续更新和补充。在开放的形式上，一般采用国家统一的门户网站形式开放数据。此外，重视建立公众的参与和反馈机制，确保用户的需求得到及时反馈，优先释放用户需求最为迫切的数据集，并对数据开放的相关进展进行评估。

从技术上来看，政府数据开放基本上都只提供数据下载服务。政府将开放的数据放在政府网站上，公众可以下载需要的数据。这些数据往往不可机读，公众更不可能通过上传到应用程序来使用这些数据。这样当数据资源比较大的时候，这些数据就变成了不可用的数据[7]。

3.2 科学数据开放

从最早推行数据资源开放的科学研究领域来看，科学数据表面上已经开放了，但实际上开放程度非常有限，主要是由政府或公共资源投资的科学研究产生的数据的开放，并且大多集中于各自领域，例如地震科学、水利科学、天文学等。在我国，主动共享科学数据的研究单位和个人还比较少，大部分的数据共享活动是通过政府投资、项目驱动的形式进行的。这些都影响了科学数据的开放共享进展和质量，目前为止，尚未形成完全开放的科学数据开放共享局面。

3.3 数据自治封闭

绝大部分数据资源还处在封闭不开放的状态，数据完全由数据拥有者自己治理，即数据自治。从20世纪90年代信息化战略开始，大部分数据是由各类计算机应用系统生产的，例如政府系统、金税工程、教务系统、超市系统、银行系统等。信息技术也只支持数据封闭，尽量保护系统数据不受外界侵害，即信息安全，例如系统设置防火墙、登录口令，制定用户级别和使用系统的功能类别等。

这些系统中的数据由系统拥有者自己管理，或者说数据由数据拥有者自己管理，称为数据自治。加之数据保持封闭不对外界开放，所以称这类数据资源管理模式为“数据自治封闭”。

在数据自治封闭模式中，使用数据的软件是事先知道的、基本内部的、数量有限的、安全可控的、隐私可控的。现有的数据资源管理技术（数据库管理系统、文件系统）和应用软件技术也只支持数据自治封闭模式，**图1为**数据自治封闭系统结构。

3.4 现有数据资源管理模式存在的问题

政府数据开放模式存在的问题是显而易见的，即数据资源稀缺性的丧失。因此，政府数据开放的基本出发点是：政府

数据是公共品，其权属属于公众，所以要向公众免费开放。然而，随着数据资源的战略性和基础性越来越显现，开放的政府数据也会被敌对国家利用，所以，政府数据开放应该是有限的，数据主权问题也越来越引起重视。事实上，国际上政府和公共数据资源仅开放了不到10%[8]，这也从另一个侧面说明政府数据开放的问题。更严重的是开放数据处于不治理或者无法治理的状态。

数据自治封闭模式的问题是数据资源只能由数据拥有者使用，没有发挥数据资源应有的价值。数据资源可以被加工再加工形成各种数据产品，服务于人们的生产和生活，从而产生巨大的价值。与数据自治封闭模式完全不同，如果将数据资源开放出来，那么使用数据的软件事先是不知道的、基本外部的、数量无限的、安全不可控的、隐私不可控的。数据开放模式示意如**图2**所示，现有的数据库管理系统软件根本无法处理数据开放的应用需求。因此，需要探索新型的数据资源管理技术。

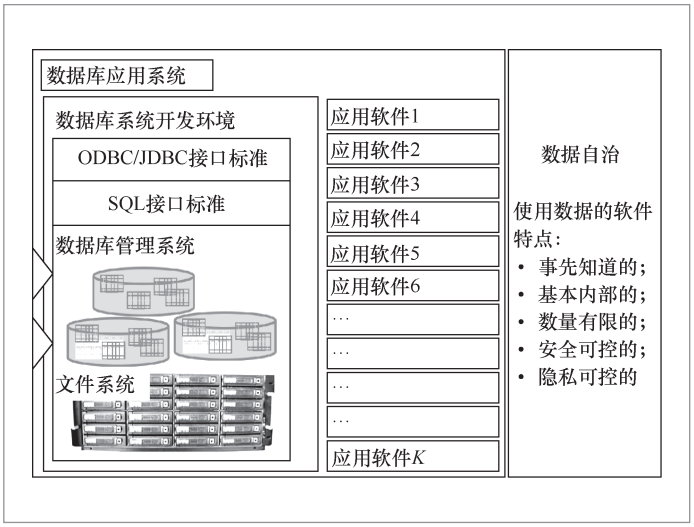

图1 数据自治封闭系统示意

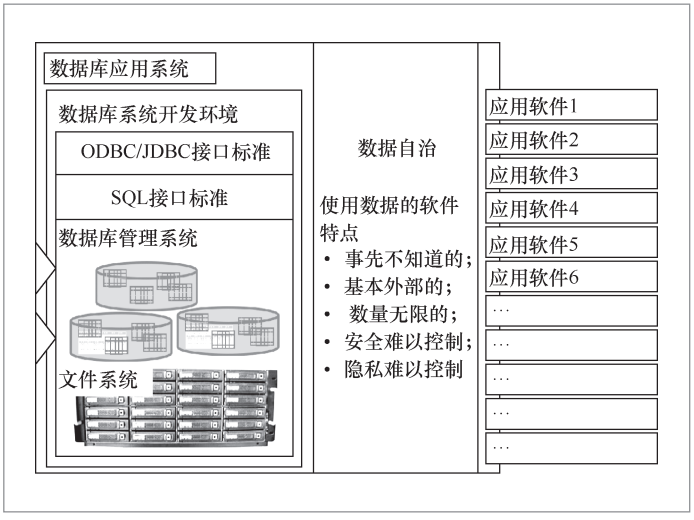

图2 数据开放模式示意

4 数据自治开放模式

数据开放是必然趋势，但需要保障在数据开放的同时又不丧失稀缺性，确保数据不流失、隐私不泄露、安全不泄密、利益得以实现，例如医疗数据的开放。医疗数据涉及相当比重和规模的隐私及敏感信息，例如患者个人信息、既往病史、就诊记录等，医生个人信息、ICD编码诊断习惯等，医院具有优势的院内制剂配方、院内诊疗规范和方案、经营财务状况等，甚至属于国家政府的涉密数据（如流行病、传染病、突发事件、重大事件等）。这直接制约了医疗数据的开放，因为没有合理有效的开放模式，医疗数据开放将增大医疗数据安全和隐私泄露的风险。为实现这一目的，数据自治开放是一种可行的方法。

数据自治开放模式是由数据拥有者管理数据，数据拥有权始终掌握在数据拥有者手里（除非自己要放弃拥有权），即数据自治；数据可以开放给指定使用者，使用者只能自己使用，不能传播数据，因此不会丧失数据的稀缺性。

为实现数据自治开放，需要开发面向数据开放的数据资源管理系统，然后将现有自治封闭系统中的数据资源重新组织到新系统中，实现数据资源的自治开放（如图3所示）。数据自治开放模式对技术提出了新挑战，数据自治开放技术要解决的问题是“如何控制数据使用者传播或滥用数据”。对应的关键技术问题如下。

● 如何做到数据既能够自治又能够开放？这需要研究面向自治开放的数据资源组织理论，即需要有新的数据模型来组织数据资源。外界能够通过这个数据模型看到有哪些数据资源，以确定是否要使用这些数据资源，系统能够承载使用者将数据上传到应用软件，根据数据模型来使用数据。

● 如何保护数据稀缺性不丧失、数据安全和隐私有保障？这需要研究面向自治开放的数据安全与隐私保护理论，确保数据使用者只能按约定使用数据，而不能传播和滥用数据。

围绕上述问题，重点研究方向包括：建立面向自治开放的数据组织模型；研究自治环境下数据使用外部软件行为管控方法；研究开放数据权益保护方法；研发面向数据自治开放的数据资源管理系统；在典型领域形成应用开放环境，开展应用。

具体包括面向数据开放的数据组织模型——数据盒模型的建模技术、数据使用的“言行一致”管控技术、数据盒加密与隐私保护技术、数据站组成管理及数据站系统的设计与实现技术等。最终的数据自治开放应用系统的结构如图4所示。

5 数据自治开放技术

5.1 数据组织模型——数据盒

如前文所述，在数据开放环境下，使用数据的软件或程序是外部的、未知的、

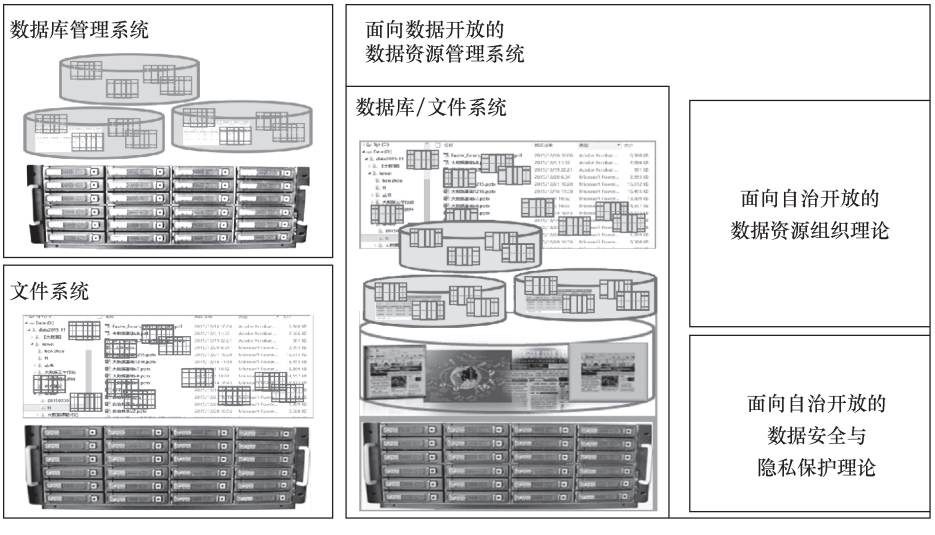

图 3 面向开放的数据资源

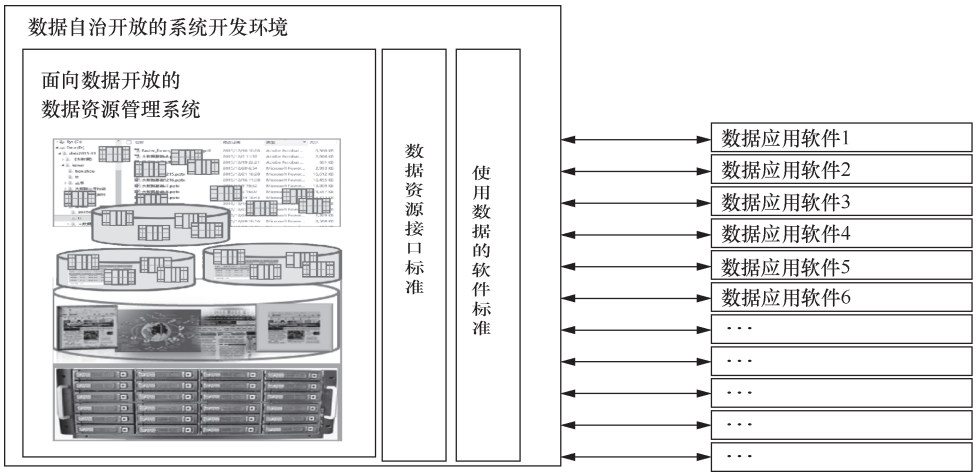

图 4 数据自治开放应用系统的结构

无限的。数据的组织既要实现开放使得用户方便使用，即数据外部可见、可理解、可编程，又要防止数据权益受到侵犯，即内部可控、可跟踪、可撤销。这需要有面向自治开放的数据组织模型，涉及以下关键技术。

（1）开放数据的基本存储单元建模技术

开放数据的基本存储单元是为数据使用者提供开放数据的基本组成单元，称为“数据盒”。自治开放模式将按照数据盒的方式向数据使用者开放数据，即呈现给用户的是一定数量的数据盒。对用户开放的数据是局部数据，不同类型数据、不同用户需求，数据开放的粒度是不同的。如何从数据属性维度（横向）和数据规模（纵向）划分数据粒度，对数据使用者使用数据、组织数据单元是一项关键技术。并且，对数据使用者开放数据的基本单元需要具有防泄露、保护权益的能力，如何将数据防泄露功能和数据权益保护机制等封装在数据单元中，是需要解决的关键技术。

（2）数据盒的形式化与计量技术

数据描述、数据操作和约束是数据盒的基本要素，数据盒的使用涉及数据盒的交、并、拼接等操作，这需要对数据盒进行形式化表示。数据盒的计量是根据数据使用者提出的要求和目标，计算使用者所需数据盒的数量和时间等，并进行定价，包括数据使用需求建模、数据需求与数据盒自适应匹配方法、数据盒的计量度量设计与度量方法、定价规则和方法等。

5.2 数据使用外部软件行为管控方法

数据自治开放环境允许数据使用者通过外部软件访问以数据盒形式存在的特定数据资源。为了保护数据利益和数据资源的可持续发展，应当对外部软件访问数据的行为进行规范化和管控。外部软件行为管控是数据自治开放中保障数据权益的重要环节。通过监控外部软件访问软件的长期行为，提取软件访问数据的行为特征，并基于这些特征抽象其

高层意图。涉及的关键技术包括以下几个方面。

（1）基于业务领域知识模型的软件行为意图建模技术

客户软件访问开放的数据资源时，应当表明其访问数据资源的高层意图。例如某客户软件声称为了追踪病症A的治疗和患者愈后情况，需要访问该病症的所有医疗数据，那么根据这一意图，对与病症A“概念相关”的数据资源的访问（可能）都是符合其意图的。这种概念相关性依赖于特定业务领域知识模型以及对开放数据资源的语义标注。在客户软件访问开放数据资源时，对其所有数据访问行为和访问过的数据资源语义进行分析，对客户软件访问数据资源的实际意图进行建模。

（2）数据使用的言行一致管控技术

在数据自治开放环境中，外部软件以黑盒方式在授权范围内对数据进行自主访问。外部软件在进入计算环境前，应当先声明其使用开放数据资源的目的，即提供其标称意图。标称意图的描述与该软件的特定业务领域密切相关，也应当表明其将采用的主要数据处理方法，作为使用数据时行为合法性的评价标准。声明了合法标称意图的软件在实施数据访问时，其行为序列应当符合其所声称的意图。根据软件行为推测得到的意图，即软件行为意图。当软件的行为意图（行）与软件标称意图（言）不一致时，即表明该软件对开放环境造成风险。为了验证外部软件行为是否符合其声明的意图，需要相应的软件行为验证技术。在隔离受控的沙箱环境中，对数据单元访问接口和环境的不同安全级别进行模拟，留存软件行为日志进行分析验证。在此基础上，在外部软件使用数据的过程中，还需要采用量化机制客观评价外部软件的行为损害数据权益的风险，通过衡量行为意图偏离标称意图的程度、行为意图对数据价值和利益相关方的影响程度、软件行为意图判断准确度等因素，综合判定该软件的行为风险等级。

5.3 开放数据权益保护方法

数据自治开放以数据盒为基本数据单元向数据使用者开放，因此数据资源稀缺性丧失和隐私泄露等问题的防范主要针对数据盒。数据盒数据被窃取、隐私数据泄露以及机密数据丢失等问题将导致数据权益受损，会降低数据拥有者开放自身数据的意愿。数据盒权益保护涉及以下关键技术。

（1）数据盒加密与隐私保护

一个数据盒可能包含照片、视频、文本和结构化数据等，数据盒的使用是外部的、未知的、无限的，传统的数据加密、数据隐私技术无法有效应用在数据盒中。

在数据盒加密方面，数据开放下的数据加密保护需要兼顾两种情形。一是在数据盒正常使用情况下，需要考虑数据盒的安全性和功能性的权衡，使得在保证数据正常高效操作的前提下最大程度地保证数据的机密性。这需要可调整的加密技术，将相应数据项进行一层或多层加密，当外部软件请求使用数据时，在保证操作（读、写、结合等）顺利执行的前提下只需要打开所需的层次，使得该层既能完成外部软件所需的操作，同时又不至于公开更内部的层次。二是即便数据盒被盗取或控制，也需保持数据盒中数据的机密性，这需要保证数据盒抗盗取和抗逆向拆解的技术。

在数据自治开放模式下，数据使用者的软件在申请使用数据盒时，需要有一个数据使用说明，说明软件使用哪些数据、以什么样的方式使用这些数据、使用的预

期结果是什么。因此数据盒的隐私保护主要包括如何判断一个数据使用说明是否涉及隐私泄露、涉及哪些隐私数据、严重程度如何。这需要研究新型的隐私认知技术。

（2）基于数据覆盖模型的数据拼图防范技术

数据拼图是指数据使用者能够通过整合多次获取的数据片段，还原数据整体。数据拼图可以由单个使用者多次获取数据片段来完成，也可由多个使用者共同合作，通过共同合作实现对数据片段的拼接。使用数据拼图技术，数据使用者可以通过非法的手段，未加授权地获取被保护的数据对象，并将其私有化。数据拼图会给数据自治开放带来实质危害，数据的使用期限、使用目的等权属将难以受到保护。而且数据使用者可以将通过数据拼图获得的数据再次传播给其他的未被授权的数据使用者，进一步造成对原数据权属的二次侵犯。首先需要构造数据使用行为的形式化描述，通过追踪分析数据痕迹，动态构造数据覆盖模型，实时检测与量化数据拼图的危害性，建立可行的防范以及预警体系，有效预防与阻止数据拼图对数据权属的侵害。

5.4 面向数据自治开放的数据资源管理系统

数据资源以数据盒的形式存放在数据站中，每个数据站配备一套数据资源管理系统，用以管理该站下的所有数据资源（数据盒）。通过数据盒虚拟化、应用装载等功能供外部软件使用数据。与传统的数据库管理系统（database management system，DBMS）相比，数据资源管理系统承担的数据管理不涉及事务处理，只有数据使用，但也不同于数据仓库，数据仓库用于数据开发利用而不是数据开放。涉及的关键技术包括以下几个方面。

（1）数据站组成与管理技术

数据资源装载在数据盒中，数据盒储备在数据站里，因此需要研究数据站的逻辑构成要素、物理形态、数据盒的组织方法与管理技术，以便能够快速定位某个数据资源的位置，包括通过数据资源元数据查找数据在哪些数据盒中，并从大量数据盒中快速定位到某一个数据盒，为用户提供数据盒，展示数据盒的内容或数据资源样本。此外，还需要研究数据盒的新增、更新、冻结（即不再对外提供使用）以及浏览、查询、校核等管理技术。

（2）数据盒虚拟化方法

数据盒的虚拟化是结合硬件虚拟化技术，为每个需要访问特定数据盒的外部软件提供一个操作托盘。各虚拟数据盒相互隔离，且对某个虚拟数据盒的更改和删除不会影响其他同源虚拟数据盒或原始的数据盒。对于数据资源管理系统而言，数据单元虚拟化技术直接关系到数据使用的安全性，即保护数据或隐私不会泄露，保障数据自治公开以及保证外部软件使用数据规范受控。需要重点突破不在物理存储上完全制作一份数据的副本的基础上，实现虚拟化的虚拟数据单元相互隔离、可用，且控制内存等资源的使用率，使整个数据站能够支撑大量外部软件，同时使用虚拟数据单元；如何在不进行数据盒物理复制的前提下提供虚拟化的数据盒，研究虚拟数据盒缓存技术、虚拟数据盒变动维护（更新、撤销等）和长操作策略等。

（3）NoSQL/Open运行库和SDK

设计NoSQL/Open（NoSQL open data language）语法规则，开发适用于

常用操作系统的NoSQL/Open运行环境和运行库，支持主流编程语言的软件开发工具包（software development kit，SDK），为外部软件实现与数据资源管理系统的互操作提供对数据站内虚拟数据盒的访问。通过NoSQL/Open，可以方便地使用数据盒。

（4）系统承载力与数据站承载力模型

数据资源管理系统和数据站也不可能管理无限多的数据，提供无限的数据访问能力。因此需要给出数据资源管理系统承载力模型，用以描述单个数据资源管理系统的极限能力、单个数据站所能承载的服务能力极限、与硬件的关系、数据站的扩展性与承载能力的关系等。

6 结束语

面对数据开放共享的战略需求，传统的面向数据自治封闭的数据管理技术无法适应数据开放的需求，急需开发面向数据开放的数据资源管理技术。本文提出了“数据自治开放”这一新型的数据资源开放模式，数据由数据拥有者在法律框架下自行确权和管理、自行制定开放规则（即数据自治），然后将数据开放给使用者，使用者没有数据治理权。数据资源稀缺性不丧失的开放才是可持续的开放，就像保护知识产权才能保护创新，才能可持续。数据自治开放模式有望成为数据开放的基本模式，是政府数据开放共享、企业及个人数据交易、国家数据主权实现的一种可行方法。

参考文献:

[1] LOUREN O R P. An analysis of open government portals: a perspective of transparency for accountability[J]. Government Information Quarterly, 2015, 32(3): 323-332.

[2] AUER S R, BIZER C, KOBILAROV G, et al. DBpedia: a nucleus for a Web of open data[C]//The 6th International Semantic Web and 2nd Asian Conference, November 11-15, 2007, Busan, Korea. Heidelberg: Springer Press, 2007: 722-735.

[3] YOZWIAK N L, SCHAFFNER S F, SABETI P C. Data sharing: make outbreak research open access[J]. Nature, 2015, 518 (7540): 477-479.

[4] 郑大庆，黄丽华，张成洪，等. 大数据治理的概念及其参考架构[J]. 研究与发展管理，2017, 29(4): 65-72.
ZHENG D Q, HUANG L H, ZHANG C H, et al. Concept and reference architecture of big data governance[J]. R&D Management, 2017: 29(4): 65-72.

[5] 朱扬勇，熊赟. 数据资源保护与开发利用[M]. 上海：上海科技文献出版社，2008: 133-137.
ZHU Y Y, XIONG Y. Protection and utilization of data resources[M]. Shanghai: Shanghai Scientific & Technical Publishers, 2008: 133-137.

[6] 黄如花，李白杨，周力虹，等. 2005—2015年国内外政府数据开放共享研究述评[J]. 情报学报，2016, 35(12): 1323-1334.
HUANG R H, LI B Y, ZHOU L H, et al. Review on the research of open and sharing government data at home and abroad in 2005-2015 [J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(12): 1323-1334.

[7] 郑磊. 开放政府数据研究：概念辨析、关键因素及其互动关系[J]. 中国行政管理，2015(11): 13-18.
ZHENG L. Study on open government data: definitions, factors and interactions [J]. Chinese Public Administration, 2015(11): 13-18.

[8] YANG T M, LO J, SHIANG J. To open or not to open? determinants of open government data[J]. Journal of Information Science, 2015, 41(5): 596-612.

作者简介

朱扬勇(1963-),男,博士,复旦大学计算机科学技术学院教授、学术委员会主任,上海市数据科学重点实验室主任。1989年起从事数据领域研究,2008年提出数据资源保护和利用,2009年发表了数据科学论文“Data explosion, data nature and dataology”,并出版专著《数据学》,对数据科学进行了系统探讨和描述。2010年创办了“International Workshop on Dataology and Data Science”,2014年和石勇、张成奇共同创办了“International Conference on Data Science”。担任第462次香山科学会议“数据科学与大数据的理论问题探索”的执行主席、“大数据技术与应用丛书”主编。目前主要研究方向为数据科学、大数据。

熊赟(1980-),女,博士,复旦大学计算机科学技术学院教授。2004年起从事数据领域方面的研究工作,作为项目负责人主持国家自然科学基金、上海市科学技术委员会发展基金以及企业合作项目。相关研究成果在国际权威期刊和会议发表论文40余篇、出版著作3本。目前主要研究方向为数据科学和大数据。

廖志成(1974-),男,复旦大学计算机科学技术学院高级工程师。1997年开始从事信息安全领域研究,2011年起从事数据领域方面的研究工作。作为主要人员参加了20余项国家自然科学基金项目、国家“863”计划项目、国家教育委员会项目、上海市科学技术委员会项目、上海市教育委员会项目以及企业合作项目。先后获得过中国高校科学技术奖二等奖1项、上海市科技进步奖二等奖2项、中国智能交通协会科学技术奖二等奖1项。发表论文9篇,出版著作7本。目前主要研究方向为数据科学、大数据、数据挖掘。

叶雅珍(1985-),女,复旦大学计算机科学技术学院助理研究员,东华大学博士生,主要研究方向为数据科学和大数据。

收稿日期: 2018-02-05

基金项目: 国家自然科学基金资助项目(No. 91546105, No. U1636207);上海市科技发展基金资助项目(No. 16JC1400801)

Foundation Items: The National Natural Science Foundation of China (No. 91546105, No. U1636207), Shanghai Science and Technology Development Fund (No. 16JC1400801)