基于城市知识体系的公共数据要素构建方法

郑宇[1,2,3]，易修文[1,2]，齐德康[1,3]，潘哲逸[1,2]
1. 京东城市（北京）数字科技有限公司，北京 100176；
2. 京东智能城市研究院，北京 100176；
3. 西南交通大学计算机与人工智能学院，四川 成都 611756

摘要
数据要素是数字经济发展的核心动能。城市公共数据的基础良好、普适性强、应用场景丰富，成为政府主导的数据要素的首选。当前数据与应用耦合，不同应用之间共享数据难，人工数据治理过程滞后、繁重低效，仅依靠自动抽取技术无法保证数据要素的精度。为此，基于人机智能协同的总体思路，提出基于城市知识体系的数据要素构建方法。首先，对大量城市业务进行解构和抽象，构建以人、地、事、物、组织5类实体，实体间关系及实体属性为核心的城市知识体系，并以这些实体、关系和属性为数据要素的原子描述，向上组合表达各种城市业务，向下形成可标准化的数据资源体系。其次，研发一套数字化控件，承载基于城市知识体系的数据要素化理论，通过灵活配置的方式开发服务于市民的各类应用，使数据在产生时就与城市知识体系关联，自动形成数据要素。最后，构建智能学习和推荐算法，更好地连接数字化控件和城市知识体系，使应用配置人员无须学习城市知识体系就能顺畅地使用数字化控件，降低了工具的使用门槛。该方法可大大提高公共数据要素产生的效率和扩大公共数据要素的规模，释放公共数据要素的价值。

关键词
数据要素；数据资源体系；城市计算；城市知识体系

中图分类号：TP399 文献标志码：A doi: 10.11959/j.issn.2096-6652.2024042

Elementarisation method for public data based on urban knowledge systems

ZHENG Yu[1,2,3], YI Xiuwen[1,2], QI Dekang[1,3], PAN Zheyi[1,2]
1. JD Intelligent Cities Technology Co., Ltd., Beijing 100176, China
2. JD Intelligent Cities Research, BDA, Beijing 100176, China
3. School of Computing and Artificial Intelligent, Southwest Jiaotong University, Chengdu 611756, China

Abstract
Data elements are the key momentum for boosting digital economy. The data generated by public services provided by governments (a.k.a. public data) is ready to be transferred into data elements, because it has been well organized in the past decade. Unfortunately, public data is strictly coupled with the systems generating them, making it difficult for different

applications to share data. The process of munul data governance is lagging, heavy and inefficient, and relying on automatic extraction method can't ensure the accuracy of data elements. To tackle these challenges, leveraging the synergy between human and machine intelligence, we propose an elementarisation method for public data based on urban knowledge system. Our method is comprised of an urban knowledge system, a set of digital controls and some machine learning algorithms. The urban knowledge system consists of entities, relationships between entities, and the properties associated with these entities and relationships, which can be used to construct different kinds of public services and form standard data representation that can be shared among different applications. Powered by the urban knowledge system, the digital controls enable governments to create different applications as public services flexibly, through a configurable way without writing any codes. Later, the information input by citizens through digital controls in these applications is transferred into data elements automatically. Finally, the machine learning algorithms assist users to use digital controls smoothly through intelligent recommendations. Our method can produce data elements automatically, efficiently and accurately, unlocking the value of data for digital economy.

Key words
data elements, data resource system, urban computing, urban knowledge system

0 引言

数据要素对于推动经济增长、优化资源配置、提升产业链现代化水平具有重要的战略意义。2023年2月，中共中央、国务院印发的《数字中国建设整体布局规划》明确指出“夯实数据资源体系，畅通数据资源大循环，推动公共数据汇聚利用，释放商业数据价值潜能”。2023年12月，国家数据局等17个部门联合印发的《“数据要素×”三年行动计划（2024—2026年）》明确指出“发挥数据要素的放大、叠加、倍增作用，构建以数据为关键要素的数字经济，是推动高质量发展的必然要求”。至此，攻坚数据要素战略课题的大幕正式拉开。

按照数据相关权益的归属，《中共中央国务院关于构建数据基础制度更好发挥数据要素作用的意见》（以下简称“数据二十条”）将数据分为公共数据、企业数据、个人数据，并针对公共数据强调加强汇聚共享和开放开发，强化统筹授权使用和管理，推进互联互通，打破“数据孤岛”。公共数据是各级党政机关、企事业单位依法履职或提供公共服务过程中产生的数据，政府对该类数据的管理与开放具有较大的推动力和掌控力。其次，相比企业数据和个人数据，公共数据具有更好的信息化基础。历经了多年的信息化建设，政府已经储备了大量的政务数据，不少城市已经开始着手数据共享和汇聚的工作。再者，公共数据的通用性和普适性强，关乎城市中的人、地、物、组织等核心元素，应用场景丰富[1]。因此，公共数据的要素化不仅有迫切需求和巨大价值，也具备更好的工作基础和实施保障。

在此背景之下，构建公共数据的数据资源体系，实现公共数据的要素化成为数据要素领域中的重中之重。然而，现有方法很难支撑这一战略目标的实现。一方面，当前的数据治理方法大部分基于事后、集中、依靠人力的处置原则，即在应用产生数据之后，通过数据治理团队来实现数据的汇聚和共享，过程耗时费力，很难规模化，数据治理的速度跟不上数据产生的速度，数据治理的成果很难在更大范围内复用和流通。另一方面，基于智能算法的自动抽取方法能在一定程度上减轻人工负担，但由于无法确保数据治理的精度，后续还

需依靠专业团队对治理结果做大量的人工校验，仍然无法实现数据要素的自动化、规模化产生。

为解决以上问题，基于人机智能协同的总体思路，提出基于城市知识体系[2]的数据要素构建方法，重点解决公共数据的要素化问题。首先，提出基于城市知识体系的数据要素构建理论，以人、地、事、物、组织5类实体，实体间关系及实体属性为数据要素的原子描述，形成数据的“元素周期表”，为纷繁复杂的数据提供有效、一致的表达基础，为公共数据的要素化提供理论基础。其次，研发一套数字化控件，承载基于城市知识体系的数据要素构建理论，灵活配置各类公共应用，为广大市民提供服务。控件产生的数据与城市知识体系自动关联，使数据在产生时完成要素化，大大降低成本并确保精度。最后，设计智能学习和推荐算法，利用人机智能协同的方式，持续提升数据要素工具的性能，使用户在无须理解知识体系的情况下，使用数字化控件来满足业务需求，大大降低工具的使用门槛。

1 公共数据要素化的目标

数据要素化是指将数据作为一种新型的生产要素，通过支持业务、数据加工分析和数据流通3种途径实现其价值[3]。公共数据要完成要素化，应实现以下3个目标。

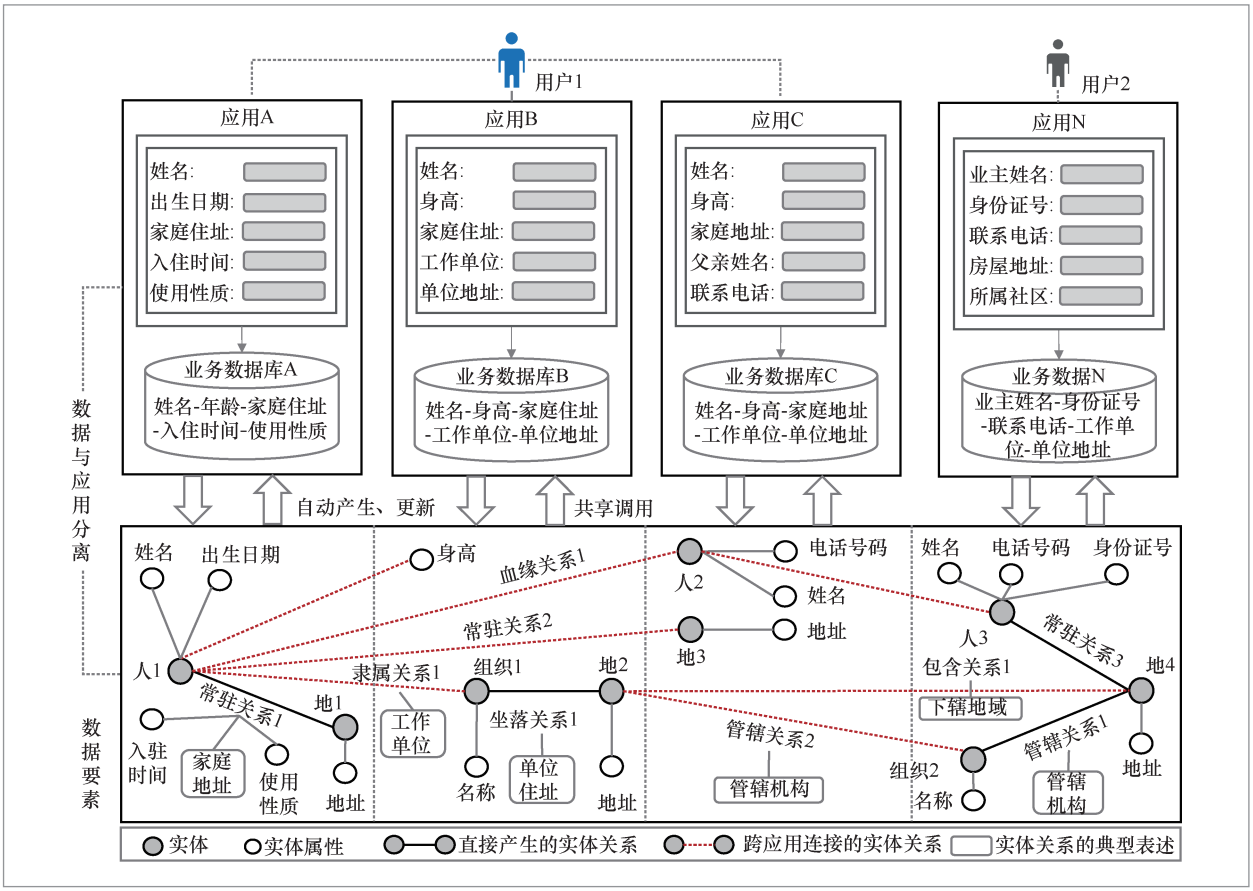

图1 公共数据要素化的目标示意图

（1）数据与应用分离[4]，以便在不同应用间精准共享

如**图1**所示，在一个应用中，业务数据虽然已通过数据库实现了与前端交互界面和后端服务程序的分离，但业务数据库的库表结构和字段含义只有该应用能理解和使用。例如，应用A存储在业务数据库A中的数据无法被其他应用调用，业务数据库A中的数据与应用A仍高度耦合。各个应用为了满足自身业务需求，需要分别重复采集数据并构建业务数据库，用户也需要在不同应用中反复填报相同信息，投入成本高、用户体验感差。

要实现数据要素化，必须从业务数据中提取反映数据本质、可共享、可理解的数据基础原件，实现数据与应用的进一步分离。如**图1**所示，用户1在应用A中填报了“姓名、出生日期、家庭住址”等信息，数据资源体系除了在业务数据库A中存储该信息，还应进一步在数据要素层存储用户1的要素信息，包括为用户1创建一个人实体节点“人1”，将“姓名、出生日期”作为“人1”实体的属性，同时建立一个地实体节点“地1”，将“家庭住址”作为“人1”和“地1”两个实体之间的“人地常驻关系”，“入驻时间”和“使用性质”是“常驻关系1”这条边上的属性。为了便于理解，本文使用了图形化的表达方式，但数据要素不一定都以知识图谱的形式存在。

当用户1通过认证登录进入另一应用B时，如果也需要填报“家庭住址”，应用A中填报的“家庭住址”要素信息应能自动准确地填入应用B界面的家庭住址栏，无须用户1再次填报（用户1可以在此基础上修改），实现数据要素在不同应用间的共享。即使应用B在创建界面时，使用的控件描述是“家庭地址”（与“家庭住址”有一字之差），数据资源体系也能识别两者本质上是同一信息，仍能使用用户1在应用A中填报的“家庭地址”。但如果应用B需要用户1的“单位地址”，之前的“家庭住址”信息则不能直接被调用，因为两者本质上是不同的含义。数据资源体系需确保数据要素在共享时准确无误。

（2）数据要素能自动产生和更新，以便形成规模效应

图1中描述的数据要素产生应自动完成，无须数据治理和人为解释。例如，当用户1在应用A中提交了信息后，“人1”和“地1”实体、它们的属性、两个实体间的关系及属性应自动在数据要素层形成。同理，用户1在应用B中填报的信息（如“工作单位、单位地址”等）也会自动形成“组织1”“地2”实体、它们的属性以及两者之间的关系（“坐落关系1”）。

当用户1在应用C中填报了新的家庭住址，数据资源体系应具备识别和更新原住址的能力，无须数据治理工程师的介入。更新过程中，可以通过与用户1的交互来确认是以下3种情况中的哪种：“原住址更名”“搬至新的住址”“增添了额外的住所”。第一种情况会修改“地1”的地址属性；第二、三种情况都会创建新的“地3”实体，并建立“人1”和“地3”实体之间的人地常驻关系“常驻关系2”。只是在第二种情况中，还需要同时修改“常驻关系1”的属性，添加该关系的“结束时间”。

（3）不同数据要素能自动连接，以便充分发掘数据价值

根据用户在不同应用中填报的信息，将人、地、物、组织相互连接，从而得到个人的多维度信息、一个组织关联的人员和地点信息，以及一个地点承载的人员和组织信息等。例如，用户1在应用B中填报的“身高”信息可以作为属性与“人1”实体相连，与应用A中填报的“姓名”“出生日期”属性共同为“人1”提供信息。用户1在应用C中通过填报“父亲姓名”创建的“人2”实

体，可以自动与“人1”相连，形成两者之间的“血缘关系1”。

不同用户在不同应用中填报的数据要素也能自动连接。例如，用户2在应用N中填报了一批数据要素，包括“人3”“地4”“组织2”3个实体及其属性、“人3”和“地4”实体间的“常驻关系3”以及“组织2”和“地4”实体间的“管辖关系1”。由于“人3”实体的“姓名、电话”属性与“人2”实体的相同，可以判断“人2”和“人3”为同一人实体，应该合并这两个节点及其属性。此外，发现“地4”实体在空间关系上包含“地2”实体，因此，“地2”也同样归“组织2”管辖，即同属于一个社区居委会管理。

将不同的数据要素连接起来可提供更深层次的价值。例如，基于零散应用中填报的数据要素，可以很快得出一个社区里不同年龄段的人数、重点关注人群、不同类型的组织数、不同使用性质的房屋分布等。数据要素的自动连接可以通过与终端用户的交互予以确认，从而确保数据要素的精度。

2 构建公共数据要素的价值

按照数据要素投入生产的途径，可将其价值释放过程划分为3个阶段，即数据支撑业务贯通、数据支撑政务智能决策、数据支撑流通对外赋能[3]。公共数据要素的价值也可以从这3个方面分析。

2.1 公共数据要素支撑业务贯通

（1）解决部门多头采集和居民重复填报问题

当前，政务及公共服务系统种类繁多，政务部门多头采集数据，市民在不同应用和场景中反复填报信息的问题严重，增加了财政负担和人力成本，降低了居民的用户体验感和参与公共事务的意愿度[5]。公共数据要素的构建能攻克这一难题，使数据在不同应用间共享、复用，减少冗余的系统建设和经费投入，并显著提升公共服务的效率和各参与方的体验。

（2）提升跨部门、跨层级的协同效率和事务处置能力

一方面，利用数据要素的共享能力，为各部门的业务开展提供所需数据，促使业务高效闭环。例如，危化品治理工作涉及应急、公安、环卫、运输等多个部门，每个部门仅拥有一部分数据，均不掌握治理过程需要的全量数据[6]。以往调取其他部门的数据需要花费很长的时间（用于整理数据、开设接口等），有时甚至无法调取，这大大增加了危化品事件的处置难度，降低了工作效率。即使花费大力气完成了某类数据的调用，也无法避免日后不断出现的新需求带来的新投入。另一方面，通过数据构建数字化协同通道，与线下的协调和推动机制相结合，使多部门在联合行动时，可以按照一条数字化主线有序、紧凑地开展工作，并及时了解进展、解决问题。

（3）赋能基层工作者，大大提升基层治理效率

首先，只有完成数据的要素化，各部门产生的数据才能自动、安全、及时地下发至基层，解决一线工作人员最需要使用数据却没有数据的问题。以往的数据制式繁杂、含义模糊、权限不清，无法直接提供给基层工作者[7]。依靠人工进行数据治理，不仅需要投入大量的人力资源，还无法保障业务的实时性以及使用数据的灵活性和安全性。其次，基层工作者也是数据要素的创造者，他们需要在一线第一时间共享自己创建的数据，而不是等待所有

数据上传至市级系统后，再经统一治理后向下分发。基层的数据共享既可提升个体的工作能力，又能促进协同的效率。

2.2 公共数据要素支撑政务智能决策

（1）支撑数据自动聚合汇总，及时把握总体态势，精准辅助决策

经过要素化的公共数据才能支撑跨层级、跨部门的数据连接和自动汇总，避免某些紧急事件使各层级、各部门临时处理数据、上报信息、合并报表，影响决策质量和效率。例如，当城市的某区域出现洪涝灾害时，政府需要第一时间掌握该区域内的居民规模，小孩、老人和伤残人士等重点监护人员的数量，房屋等建筑的数量和分布，医院及学校等机构的组织情况，这些信息在各种应用中已经产生，但分散在各个业务的数据库中，只有经过要素化才能相互连接起来，从而辅助政府人员及时、精准地制定疏散和救助方案。

（2）支撑多维、异构信息的连接，挖掘深层次的知识

通过聚类、分类、回归等数据挖掘方法，分析公共数据要素中蕴含的深层次的关系与规律，突破人们认知的局限，辅助智能决策。例如，为了实现对重大传染性疾病的精准防控，发现病毒感染者后，需要对他们接触过的人员、去过的地点、所属的组织进行排查，并进一步分析密切接触人员的相关地点和组织信息。这需要基于零散数据构建一个人员、地点和组织的图谱，通过对图谱的分析得出这些人员、地点和组织的风险等级，然后按优先级精准处置风险对象，利用有限的资源及时阻止疫情扩散，同时避免大范围封控造成的损失。由于人员、地点和组织三者之间相互影响，即去过更多风险地点的人员风险更高、越多风险人员到访过的地点风险越高、包含更多风险人员的单位组织风险更高、在风险越高的组织里工作的人员风险越高等，需要利用数据挖掘算法计算这些对象的风险等级，从而辅助决策，而人类直觉和简单的数据统计无法实现这一目标[8]。

2.3 公共数据要素支撑流通对外赋能

（1）解决数据交易中的数据资源互认问题，支撑交易线上化和规模化

数据交易要形成规模，需要摆脱对线下人工撮合的依赖，实现线上自动交易。要实现这一目标，首先要解决不同机构提供的数据资源间的互认和识别问题。由于这些数据都是在各机构的业务系统中产生的，缺乏统一的数据资源标准，在交易前需要经过数据治理，在交易过程中需要附加人工解释和理解的成本，无法实现交易过程的自动化。公共数据的要素化将解决数据交易环节中的这一难题，为数据交易的线上化和自动化打下基础。

（2）促进数据要素在政府体系外的大循环，提升数据资源利用率

公共数据要素不仅可以用于公共事业，也能帮助企业提升业务质量，增加经济效益。例如，政府侧的失信人员和企业名单可以提升银行信贷业务的风控质量，降低资产的不良率；政府登记的房产信息能帮助地产公司为居民提供合适的优惠方案；企业的纳税和社保缴纳情况能帮助从业人员甄别和选择优秀企业。如果这些信息都零散发布，企业查找、使用和关联数据的成本很高，数据的安全性也缺乏统一的保障。公共数据的要素化将大大降低企业和社会组织使用公共数据的门槛，使数据资源能够在更大范围内被循环利用，在公共服务之外产生更大的经济价值和社会效益。

3 公共数据要素构建的现状和挑战

数据要素的3次价值释放过程，涉及3类不同的技术：一次价值的释放过程主要基于大数据管理技术[9]；二次价值的释放过程涉及大数据治理、处理和分析技术[10-11]；三次价值的释放过程需要大数据流通、交易和隐私保护技术[12-13]。参照数据要素的3次价值释放过程，可将构建公共数据要素的相关工作分为3类：公共数据管理、公共数据治理和分析、公共数据交易流通。

同时，按照数据结构维度的划分方法，公共数据大致分为以下3类：以电子政务表格为代表的结构化数据、以图像、语音和文本为代表的非结构化数据、以地理信息和物联网数据为代表的时空数据[14]。经过多年的发展，非结构化数据在业界已经形成了通用的制式，如常用的视频文件格式有AVI、MP4、WMV等，常用的音频格式有MP3、WAV等，常用的文本文件格式有txt、Word、WPS、PDF、XML等。各类系统能够自动产生非结构化数据，数据制式之间相互转化的标准和工具也非常成熟，非结构化数据可以很方便地在不同系统间被调用和共享，其要素价值得以释放。另外，时空数据的要素化也取得了一定的进展。基于数据结构的时空属性和数据关联的时空属性两个维度，Zheng[15]提出了6类时空数据模型，并搭建了实用的京东城市时空数据引擎（JD urban spatio-temperal data engine，JUST），为时空数据的一致性表达和数据管理系统的可扩展性提供了理论基础。而对于结构化公共数据的要素化问题，相关研究工作非常少，这部分数据的要素价值亟待释放。因此，本节针对结构化公共数据的要素化问题，从公共数据管理技术、治理和分析技术以及流通技术3个方面展开现状和挑战分析。

3.1 公共数据管理技术

公共数据的管理重点是解决各级党政机关、企事业单位依法履职或提供公共服务过程中数据的接入、存储、修改和查询问题，通过信息系统建设和应用开发，实现业务的数据化。同时，每个系统都会构建服务于自己的数据库或者存储体系，大部分数据已经存储在政务云上。当前该领域的工作主要集中在优化存储效率、提升查询和访问速度以及增强数据存储的安全性和稳定性方面，涉及的大数据管理技术主要包括分布式文件存储系统（如HDFS）、关系型数据库技术（如Oracle、MySQL、Microsoft SQL Server等），以及非关系型数据库技术（如HBase数据库存储系统和Hive数据仓库系统等）。

数据贯穿了公共服务的业务流程以及线下与线上的边界，提高了工作效率，储备的数据也为后续要素价值释放打下了基础。但数据存储体系产生的数据与应用高度耦合，库表的结构、字段的命名和数据的血缘高度依赖应用的业务逻辑。由于没有统一的数据构建标准和工具，一个信息系统产生的数据无法被其他应用调用。因此，在需要跨部门、跨系统使用数据时，不得不让技术人员将一个系统中的数据手工导出，处理后再发给需求部门的技术人员，后者再将其导入另一个系统中使用，这个过程代价巨大、时效性差。

为了解决这一问题，很多地方政府开始构建共享交换平台[16]。其基本原理是各部门将自己拥有的数据编目提供出来，放在一个共享平台上，形成一个大的数据资源目录，但数据仍存储在各部门的存储体系中。当某部门需要调用数据时，可以从该目录上查询到相关部门的具体数据条目，

并发起调用数据的请求。如果存在该数据的数据调用接口，数据供需方达成共识后，需求方可以通过该接口自动调用数据。如果该数据是首次被请求，数据拥有方需要开发数据调用接口，并将接口注册到共享交换平台以供需求方使用。共享交换平台通过数据资源目录和数据调用接口缓解了一部分调用数据的压力，减少了一些手工调取数据的操作，但仍然存在大量的沟通协商和临时开发的工作，无法满足大规模调用数据的需求，数据调用的时效性仍然较差。

3.2 公共数据治理和分析技术

为了进一步利用不同应用系统产生的数据，数据治理技术应运而生[17]。其基本原理是利用各种大数据工具将各业务系统中的数据汇聚到一个集中的数据平台中[18-19]，通过设计有一定共性的元数据和库表结构，按照新的应用需求对数据进行分层管理、逐级聚合。该设计提高了数据的利用率，解决了一个项目中上层应用需要重复调取原始系统中的数据的问题。

数据治理过程涉及数据接入处理技术和数据分析处理技术。其中，数据接入处理技术包括针对流数据接入的Kafka技术、建立关系型数据库与HBase之间联系的Sqoop技术、收集日志的Flume技术、批量传输文件的FTP（file transfer protocal）技术以及获取互联网特定目标地点数据的网络爬虫技术等。数据分析处理技术包括Spark、Storm和Flink等流处理技术。

然而，当前的数据治理技术仍面临以下挑战。首先，在应用产成数据后，数据治理工程师需要先理解每个数据字段背后的含义和生成逻辑，再按照数据应用的需求对原始数据进行清洗、转换、分析，过程耗时费力，很难规模化；其次，数据治理过程严重依赖于工作人员的个人能力和工作习惯，不同的人对数据的命名规则和组织方式不同，不同团队使用的治理工具也不同，导致数据治理的成果很难在更大范围内复用和流通，被治理的数据之间的联系也无法自动被发掘，只是变成了更大的“数据孤岛”；最后，当前的数据治理与数据应用耦合紧密，数据如何治理取决于新的应用打算如何使用这些数据。因此，只要有新的应用数据需求，数据治理的工作就永远无法停止，很多智慧城市项目近一半的精力耗费在数据治理上。

要解决以上难题，数据治理工作需要突破以下两个瓶颈。

（1）制定标准的数据基础元件

标准化的数据基础元件是数据要素化的基础和保障，它可以促进数据的互联互通、互认互用，降低数据的流通成本，提升数据的流通效率。为了推动数据基础元件的标准化，《数据元件的结构要求》[20]、《数据元件安全审核要求》[21]两项团体标准为建立数据元件的设计、开发、管理和维护机制提供了参考。国家标准《智慧城市 领域知识模型 核心概念模型》[22]中提出了智慧城市领域的核心概念、模型组成、核心理念以及核心概念之间的关系。此外，对于特定业务的数据标准，不同城市、不同行业对其均有规定，其数量较多，例如国家标准《智慧城市 数据融合 第5部分：市政基础设施数据元素》[23]中规定了智慧城市市政基础设施的数据元素。但数据元素到底包含哪些具体内容、如何融入实际的公共业务中并完成数据要素化的3个目标，还需要进一步探索。此外，标准要能成功应用，除了凝聚共识、大力推行，还需要高效、简洁的自动化的工具来承载标准，降低标准的推广代价。当前，从标准到智能化工具的路径仍不清晰，可用的公共数据要素化的工具仍然缺失。

（2）建立业务表单与标准化数据基础元件之间的映射

目前主要有以下两类方法。

第一类方法是基于人工的数据治理。例如国家标准《智慧城市 数据融合 第3部分：数据采集规范》要求对原始数据进行清洗、转换、分析等[24]。人工操作首先需要综合分析元数据的信息，然后观察实际数据，进而判断数据类型，理解数据背后蕴含的语义信息[10]。之后，再按照数据应用的需求，将业务数据转换到某预先设计的数据基础元件中。这背后的一项关键技术是数据整理，包括结构化处理、数据质量评估、数据清洗、数据规划法、数据融合与摘要等[25]。数据治理技术对数据的使用和共享做出了不小的贡献，在很多智慧城市项目中被广泛使用。但基于人工处理的方式耗时费力，且依赖于个人的工作能力和工作习惯，不利于数据要素的规模化形成和大范围流通，需要进一步更新。

第二类方法主要基于机器学习的自动抽取技术。现有研究通过综合数据源的统计信息、语言信息[26]或综合数据元属性及数据上下文[27]，使用机器学习的方法对数据进行标注。例如，数据库领域的列语义识别技术[28-29]，可实现列名和标准元数据的语义对齐；知识图谱领域的实体关系抽取技术[30]，面向预先设定的本体结构[31]，通过语义解析将不同数据整合到知识图谱中[32]。自动抽取技术的精度受数据质量和数据复杂度的影响较大，尤其是当表单中具有多个实体、多个关系且元数据内容繁多时，自动抽取的精度将大打折扣。虽然这些自动化方法能在一定程度上减轻人工负担，但由于无法确保数据治理的精度，后续还需要对治理结果进行大量的人工校验。

同时，如果数据在产生时不能要素化，即便有统一的标准和自动化的工具，后续的数据治理代价仍然非常大，数据治理很难跟上数据产生的速度，后向的集中式治理也会成为新的瓶颈，数据要素化的道路仍然艰难。

3.3 公共数据流通技术

要实现数据在更大范围内的流通，需要具备以下4个方面的条件：形成互通互认的数据资源体系、搭建隐私计算技术平台、健全数据流通交易的法律法规、打造良性健康的商业模式。

首先，供需方按照统一的标准形成互通互认的数据资源体系，这是数据开始流通交易的起始条件。尤其是针对已经产生的数据，如何利用智能工具实现自动或半自动的数据转化、调取和使用，是非常重要的。

其次，为了确保数据安全和用户隐私，以及实现数据多种权属的分离，需要发展跨域数据管理和学习技术[33]。例如，基于联邦学习和多方安全计算的隐私保护计算技术[13]，可在数据不离开原始数据库的情况下，使算法进入各数据存储体系进行计算，实现多方数据知识的融合，而不泄露任何原始数据的信息。当前，国内几家大数据交易所已经开始搭建包含隐私计算技术的平台，逐步将数据的交易从线下转到线上。

再者，健全数据流通交易的法律法规。根据数据的来源和数据的生成特征，分别界定数据生产、流通、使用过程中各参与方享有的合法权利，建立数据资源持有权、数据加工使用权、数据产品经营权等分置的产权运行机制[34]。法律法规的健全应与平台系统的搭建同步开展、相互促进，利用技术平台承载法律法规，在实现过程中细化法律法规。目前，国家已经公布了宏观的指导建议，具体的法规细则还需随数据交易的实践逐步深化。

最后，数据要素的规模化流通需要打

造良性健康的商业模式，满足大数据交易所、交易服务代理机构、数据提供者和需求方的利益，具体包括针对数据应用的报价机制、交易撮合机制、数据交易的抽成机制、代理数据交易服务的分佣机制等，这些工作需要依靠技术来实现，可参考股票交易市场的相关机制。

4 公共数据要素构建方法

4.1 总体思路

为了满足公共数据要素化的迫切需求，基于人机智能协同的总体思路，提出基于公共城市知识体系的数据要素构建方案。一方面借助人类智能确保公共数据要素构建的准确性，另一方面借助机器智能提升公共数据要素构建的自动化水平。

首先，如**图2**①–1和①–2所示，基于专家智能构建城市知识体系，形成公共数据要素理论基础。通过对大量城市公共业务的解构和抽象，获得以人、地、事、物、组织5类实体，70多种实体间关系及600多项属性为核心的城市知识体系。以城市知识体系中的实体、关系和属性为原子描述，向上组合表达各类城市业务，向下作为公共数据的元件（类似数据的“元素周期表”），形成基于城市知识体系的数据要素理论。该理论为纷繁复杂的数据提供有效、一致的表达基础，使数据与应用分离，不同的数据可以连接，产生标准化的数据资源体系，有助于后续数据高效、便捷地交易和流通。城市知识体系的内容高度抽象、数量精简，具有很强的概括性和通用性，便于理解和使用，大大增强了后续构建自动化系统的可行性。

其次，研发一套数字化控件作为数据要素的构建工具，承载基于城市知识体系的数据要素理论。数据要素的构建工具为各类实体属性设计了专有控件，业务人员

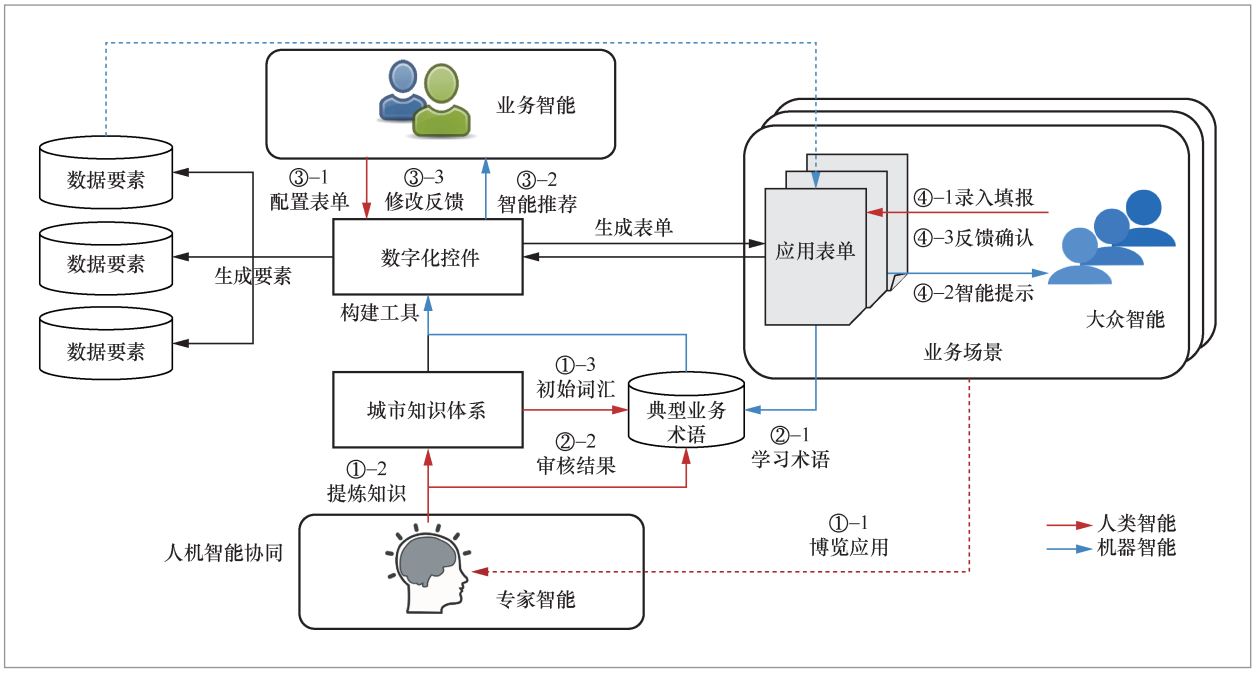

图 2 基于城市知识体系的公共数据要素构建方法的总体思路

（通常为政务部门的信息管理员）可结合自身的业务知识，灵活配置各类公共应用的表单，为广大市民提供服务。如**图2**中红色箭头③所示，在配置过程中，业务人员根据业务逻辑将相应的控件分别添加到应用界面，通过控件的添加顺序和空间包含关系来体现实体间的关系，并可根据场景需要修改控件描述（如将人的“姓名”改为“曾用名”）。

如**图2**中红色箭头④所示，基于数字化控件配置的表单生成为居民提供在不同场景中填报信息的界面，通过控件录入的数据与城市知识体系相关联，让数据在产生时就自动完成要素化。用户在不同应用中通过数字化控件填报的数据要素也能自动连接。当用户填报的要素信息发生变化时（如家庭地址发生变更），数字化控件会通过智能提示向用户寻求反馈，以确保数据的精准性，如图④-3所示。数字化控件将数据要素的产生方式从事后、集中式的繁重治理模式转变为前置、分布式的自动产生方式。

图2中的红色箭头体现了公共数据要素化方案中利用的3种人类智能。首先，通过专家智能凝练城市知识体系，并基于此研发了数字化控件；其次，基于数字化控件，融入政务部门信息管理员的业务智能，产生应用表单和界面；最后，利用表单获取居民的大众智能，形成数据要素。

为了进一步提高工作效率和用户体验，降低数字化控件的使用门槛，除了人类智能，还需要融合**图2**中蓝色箭头表示的机器智能。

在使用数字化控件配置应用的过程中，尤其是表达实体间的关系时，由于配置人员熟悉业务场景但并不一定能理解城市知识体系，数字化控件在前台界面呈现的是各种常用业务术语（如“家庭住址”），而系统在后台需要基于城市知识体系抽象、共性的表达来建立数据的要素标签以及数据之间的要素关系（如“人-地常驻关系”）。因此，需要建立业务术语和城市知识体系（尤其是实体间关系）之间的映射。

如**图2**中①-3所示，专家可以通过经验总结出一部分常用的映射作为初始数据，例如，“人-地常驻关系”的业务表达有“家庭住址”“临时住址”“居住酒店”“户籍地址”等，但业务术语的空间很大，且不断会有新的业务术语产生，很难单纯依靠人的经验来完成。此外，业务术语之间也有很大的相似性和冗余性，如“临时住址”和“暂住地址”其实本质上是相同含义，既没有必要也不能把一类实体间关系对应的所有业务术语都涵盖进来，否则会给用户制造巨大的阅读障碍和选择代价，降低工具的有效性。为此，如**图2**中②-1所示，需要智能算法从外部业务表单中不断学习并归纳出常用的、典型的新增业务术语作为补充，建立城市知识体系与不同的业务场景间的关联。如**图2**中②-2所示，专家也可以定期审核算法产生的典型业务术语，用较小的人力代价换取高质量的结果。

此外，在使用数字化控件配置应用的过程中，即便已经建立了城市知识体系和常用的、典型的业务术语之间的映射，但业务术语的数量相比交互界面上有限的呈现空间仍显巨大。如果直接全量展现，用户需要从众多业务表达中遴选，不相关的信息将变成阅读障碍，会增加配置难度、降低工作效率，甚至可能导致无法配置应用。

为此，如**图2**中蓝色箭头③-2所示，需要设计控件标题智能推荐算法，基于用户配置当前界面的过程和以往的配置经历，理解用户期望表达的意图（如实体间关系），智能推荐用户希望使用的业务术语的集合，并按可能性进行组合排序，显著提升

了用户的配置效率和体验，降低了数字化控件的使用门槛。用户可以对选择的推荐术语进行进一步修改，修改的日志也会进一步反馈给业务术语学习算法，不断提升算法的性能和术语映射的质量。

业务术语学习算法和控件标题智能推荐算法利用机器智能，让数字化控件成为用户和城市知识体系的智能连接器，在无须用户知晓和学习城市知识体系的情况下，将纷繁复杂的业务数据转换为标准统一的数据要素。

数字化控件将用户通过应用填报的信息与已经形成的个人信息要素进行比对，如发现新增信息（如不同的家庭住址），数字化控件会提示用户选择更新原信息或者添加新的个人信息。此后，数字化控件可以在新的应用中使用新信息，为用户提供更加便捷、优质的服务，但该用户在以往应用中填报的历史信息（即**图1**中的业务数据）并不会改变，也不应该改变。因为用户在一项服务中填报的信息与该服务提供的时间有关，这些信息是该用户当时状态的真实写照。

4.2 城市知识体系

基于大量智慧城市系统建设、应用示范和产业实践的经验，笔者逐步发现城市业务场景可以拆解成人、地、事、物和组织实体的知识化表达，如**图3**所示，在重点人群信息的采集应用中，居委会为社区居民服务这一场景可抽象为社工（人）和居委会（组织）的隶属关系、居委会（组织）对某房子（地）的管辖关系、居民（人）和某房子（地）的常驻关系以及对应实体及关系背后的属性（如人的姓名、性别、年龄等属性，常驻关系的入驻时间、使用性质等属性）。在安全隐患排查中，城管对井盖的维修场景可抽象为运维工（人）和井盖（物）的运维关系、运维工（人）和城管（组织）的隶属关系以及对应实体及关系的属性。

通过对大量城市业务的解构和抽象，

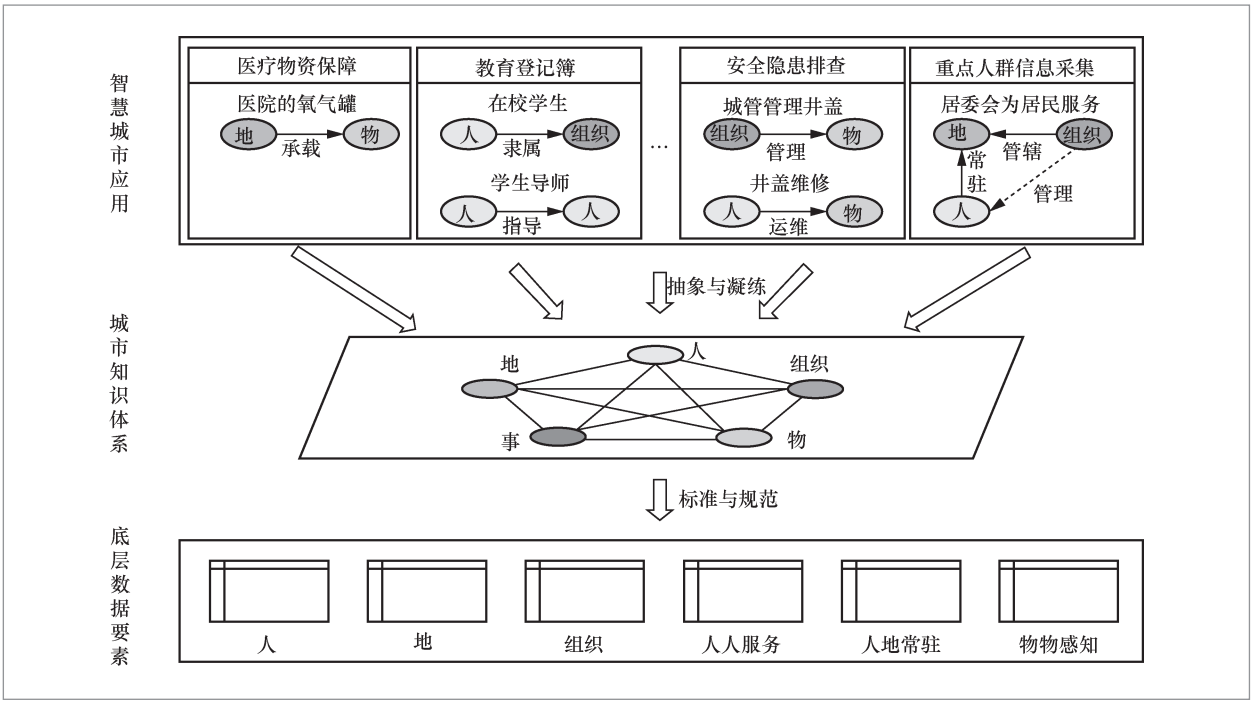

图 3 城市知识体系的底层逻辑

依靠专家智能凝练出不同场景中的共性实体、实体间关系和属性。这些共性知识揭示了不断演变的城市业态的本质规律和底层逻辑，为各类业务系统的开发和数据组织提供了一致的原子描述。基于这些原子描述构建的系统所产生的业务数据，自然被赋予了共性知识的标签，从而能做到互联互通、相互理解。以往也有一些基于本体论构建的知识体系，对城市中的一些实体及属性进行抽象和定义，但缺乏对实体间关系的体系化凝练，暂不能满足不同业务的表达需求，也无法使不同数据要素相互连接。

如**图4**所示，城市知识体系的内容包括人、地、事、物、组织五大类实体，76类实体间关系和600余项实体及实体关系属性。实体间关系由同类关系、相互关系组成。例如，人–地关系中包括常驻、访问、途经等关系，组织–组织关系中包括从属、协助、竞争等关系。每类实体和实体间的关系都有相关的属性，例如，物的属性包括名称、重量等，人地常驻关系的属性包括入驻时间、结束时间、常驻类型等。人地常驻关系是对家庭居住、旅游度假、工作派驻、医学隔离、健康疗养等10余种业务场景的抽象化表达，其本质都是一个人实体在一个地实体上驻留了一段时间。

城市知识体系中的“事”包含时间、人、地、组织、物的属性，（事件中）实体间关系，事件的结果等，如“2024年5月10日15点，O组织的工作人员B对住在A社区的独居老人X进行了慰问关怀，B向X赠送了Y物品，帮助X修理了Z设施，并获悉老人对上门医护服务有需求，本次慰问获得了老人X的高度评价”。事件之所以存在，是因为一些事务的过程与结果一样重要，如关注过程，就需要将其定义为事件。在上述示例中，如果只记录了工作人员B慰问了老人X这个结果（可用“人–人服务关系”来表达），其价值将大打折扣，工作人员B的

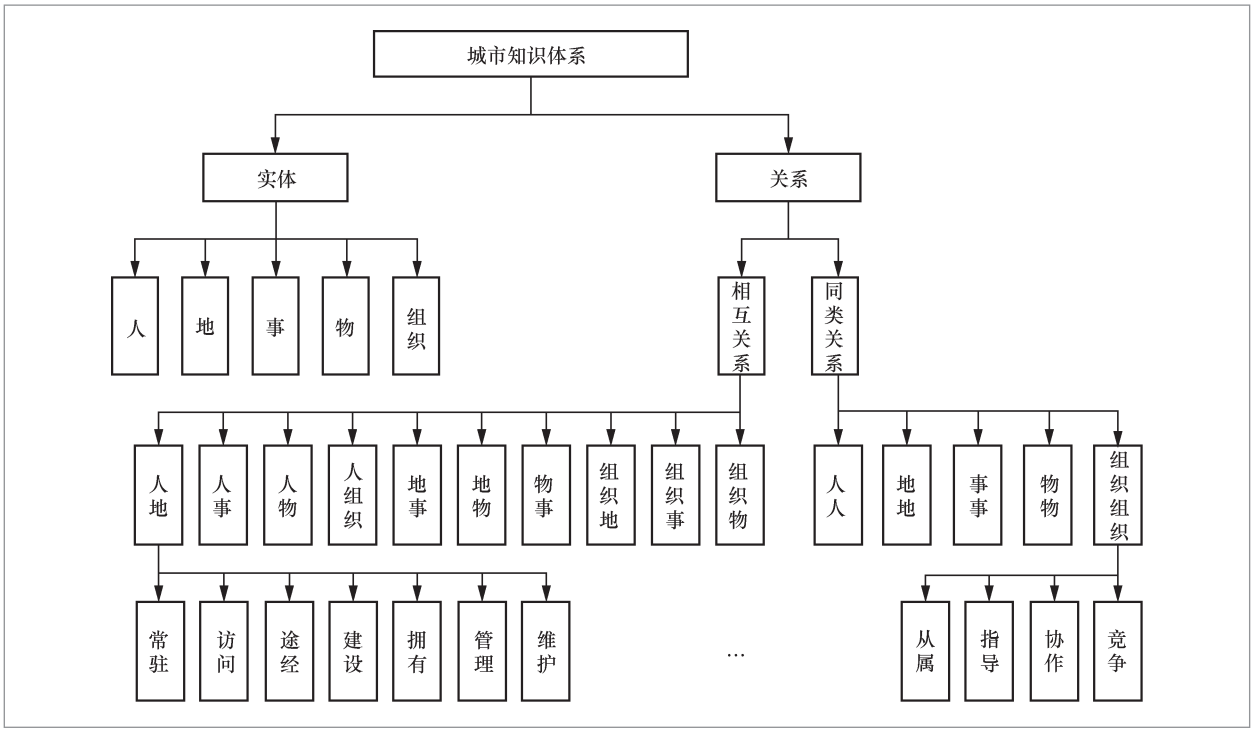

图 4 城市知识体系的内容示意图

慰问质量如何、老人X需要什么样的关怀、是否达到了组织的预期、是否获得了老人X的认可均不得而知。因此，要将其定义为一个事件来进行管理。同样，一例传染疾病的检测也是一个事件，该病例的采样时间、机构、人员和方式、检测时间、机构、方式和结果参数等都会决定后续将采用何种不同的应对措施。

这些实体、关系和属性具有高度的概括性和普适性，可以组合起来描述不同的业务场景，便于用户理解和使用，同时，其高度精练、数量有限，便于系统构建和开发调用。城市知识体系的详细内容可参看文献[2]，本文提出的数据要素化方法可以看作城市知识体系针对结构化数据的一个具体应用，是其理论框架的一个实例化的过程。由于城市公共业务的核心宗旨是“服务和管理城市中的居民和机构”，文献[7]从哲学层面论证了城市知识体系能精准地表达宗旨中的关键要素（城市、居民、机构、服务、管理），能满足公共服务业务的需求。当然，城市知识体系也需要在使用过程中不断优化内容、校正描述、补充遗漏。

4.3 数字化控件

为了实现搭建理论到应用，需要一套数字化控件来承载基于城市知识体系的数据要素理论。数字化控件通过应用界面配置、用户交互反馈、数据要素转化，衔接专家学者、业务人员和居民用户3类人群，融合专家智能、业务智能、大众智能和机器智能四大智能，完成公共数据的要素化。

如**图5**所示，数字化控件由前台界面、后台系统和数据层构成。

如**图5**右侧白色区域所示，数字化控件在前台界面为5类实体的属性设计了专有控件（如人的属性p_1、物的属性p_x等），并以知识体系提练的共性名称为控件的初始描述（如控件p_1的初始描述为“姓名”、控件p_x的初始描述为“型号”）。这些控件可以组合起来，在左侧形成“××应用”的界面，并可以根据场景的需要修改控件的初始描述（如将“姓名”改为“曾用名”、“型号”改为“设备型号”）。

如**图5**中部蓝色框架所示，后台系统提供城市知识体系的内容及其关联的业务术

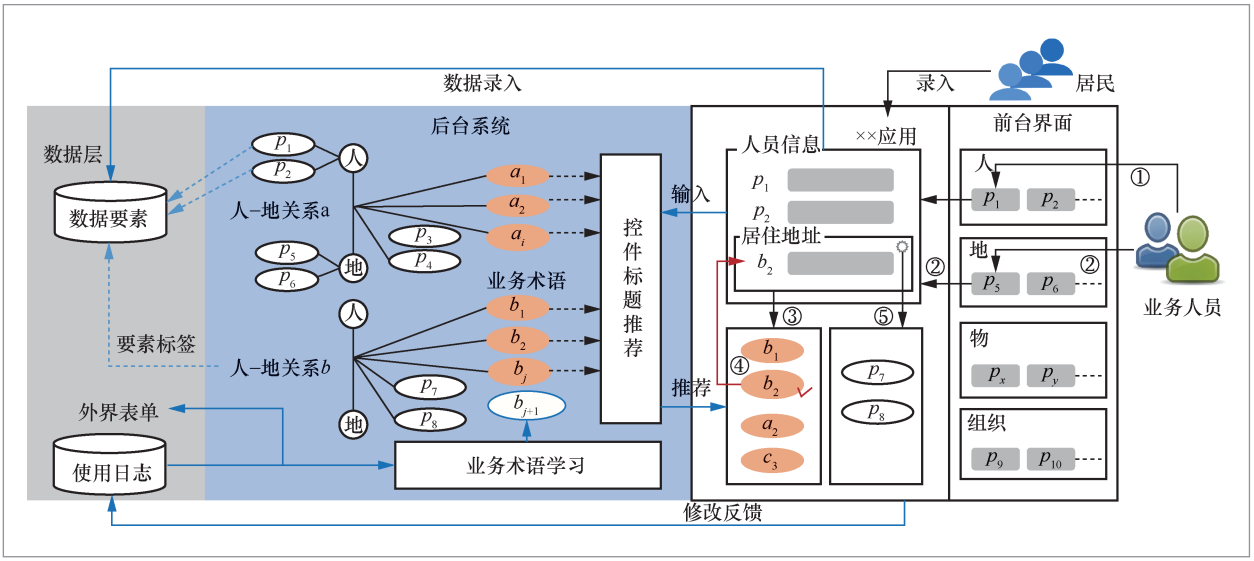

图 5 数字化控件的构成和工作原理

语，例如人–地关系a拥有p_3和p_4两个属性以及a_1、a_2、□□□、a_i多个业务表述，并持续从外部业务场景和数字化控件的使用日志中学习并增补常用的、典型的业务术语（如b_{j+1}）。后台系统在应用配置过程中分析业务人员的配置意图，推荐实体关系对应的业务术语，实现业务逻辑和知识体系的映射，降低用户使用数字化控件的门槛。在应用配置完成后，将居民在应用界面输入的信息解析为带有城市知识体系标签的数据要素，送至数据层进行存储。

如**图5**左侧灰色框架所示，数据层记录配置界面过程中的行为数据（如业务人员添加控件的顺序以及修改控件描述的日志），供后台系统在后续学习业务术语和优化推荐算法时使用。同时，存储后台系统解析过的居民输入数据，并基于城市知识体系连接不同应用产生的数据要素，为后续居民再次使用自己的信息提供准确的数据源。

图5①–⑤所示为使用数字化控件配置应用界面的具体过程。

①业务人员利用人实体控件来构建关于人的信息，并通过一个信息框展示同类实体的多个属性控件。例如，将控件p_1拖入应用界面中，在拖入首个控件时，数字化控件会根据控件类型自动生成相应的信息框（此处为人的信息框）。之后，再将控件p_2及其他所需要的人实体属性控件逐一拖入该信息框中。

②数字化控件通过控件在应用界面上的添加顺序和空间包含关系来体现实体间的关系。例如，在人的信息框中拖入一个地的控件p_5（初始描述为“地址”）。由于控件p_5是该应用界面中的首个地实体的控件，数字化控件会自动生成一个地的信息框来展示控件p_5。后续可以在这个信息框中继续添加更多实体的属性控件以及人地关系属性的控件。这些人、地的信息框及其标题只用于体现实体间的关系，辅助应用完成配置，不会呈现在最终的用户界面上。

③在地信息框生成的同时，系统会根据前面已经配置的内容（如人的信息p_1、p_2），利用控件标题推荐算法，智能地提示与“人地关系”相关的业务术语（如暂住地址b_1、家庭住址b_2、途经地点a_2、管辖地域c_3等）以供业务人员选择。

④业务人员通过选择业务术语来告知系统他们期望表达的实体间关系，而无须知晓城市知识体系。例如，选择“家庭住址b_2”，系统在后台就会自动选择“人–地关系b”，即常驻关系，从而建立控件p_5与人的信息框之间的人地常驻关系。用户也可以进一步修改业务术语作为控件的描述（如将“家庭住址”改为“居住地址”），以满足多样的业务需求。修改后的术语名称不仅会替代控件p_5的控件标题，也会替换地信息框上的提示名称。

⑤点击实体关系的设置按钮可添加更多的人地常驻关系的属性，如“入驻时间p_7”“使用性质p_8”等。这些属性都会添加到“居住地址”信息框中控件p_5的下方，应用配置人员同样可以修改这些属性的初始描述。

如果希望生成事件，可在配置好的应用界面的基础上，进一步利用事件模板设定事件的具体内容。将应用界面中已经添加的属性控件按照希望的事件描述顺序，添加至事件模板的相应位置。之后，在应用界面的信息填报完成后，数字化控件就会按照模板设定的格式，将这些控件获取的信息衔接起来，自动生成事件。以前文介绍的老人关怀事件为例，在事件模板中可按顺序选择时间控件、工作人员姓名、“人–组织隶属关系”的类型属性、组织名称、“人–人服务关系”的类型属性、老人姓名、“人–地常驻关系”的类型属性、地点名称、“人–人赠与关系”的类型属性、物品名称等，从而自动生成具有语义的事件。

通过以上方式，政务部门的信息管理员可以基于场景需求和自己的业务知识，利用数字化控件灵活配置各类公共应用，为广大市民提供服务，但又不需要学习和理解城市知识体系。数字化控件能将数据要素的产生方式从事后、集中式的繁重治理方式转变为前置、分布式的自动产生方式。

5 应用示范

除了上述理论和工具，要实现公共数据要素的规模化产生，还需要关键业务的牵引，在实际的公共服务中将理论、工具应用起来。经过多年的实践发现，基层治理业务可作为产生公共数据要素的抓手，原因如下。

- 基层作为产生和使用公共数据的最前沿，数据质量和鲜活度高。由于居民主要生活在社区等基层单元，此处民生需求集中且旺盛，居民有寻求服务和提报信息的意愿。同时，政务基层工作者与居民互动频繁，信息互通的需求强烈，基层单位也是感知居民信息变化的最前沿。

- 现有的信息化手段无法满足基层治理业务的需求，存在“真空地带”。基层治理业务的特性和示例如图6所示，基层治理业务具有种类繁多、分化细碎，临时突发、短期紧急，地域不同、处置各异三大特点。统一固化的信息化系统无法满足差异化的需求，也跟不上业务的快速变化。通过定制化开发的方式研发大量细碎的系统，存在成本高、周期长和体验差的问题。现有的电子政务服务和政务热线12345均无法满足基层治理业务的需求[7]。

因此，依靠基层治理业务牵引，以业务中政民信息互通为抓手，利用本文提出的数据要素理论和数字化控件技术，通过灵活配置的方式高效地构建各类公共服务，攻克基层治理过程中的困难，满足业务需求。文献[7]提出了政民互通的信息通道，提炼了基层治理业务中的五大共性原子能力，即地理层级、实体台账、协同任务、信息互通和权限体系，并构建了政民互通平台。将数字化控件融入该政民互通平台，支持五大原子能力的表单配置过程，产生面向居民的应用界面，向上支撑基层治理业务，向下形成公共数据要素。该方法已经在北京市的多个街道投入使用，配置公共服务和基层业务数百项，服务数百万居民，自动形成关于人、地、事、物、组织的数据要素过亿条，这些数据可以在不同的公共服务中被共享调用，无须人工进行数据治理，居民也无须在不同的应用中重复填报自己的个人信息，既为基层治理业务降本、增效、减负，也大大提升了数据要素的产生效率，增大了数据要素的产生规模，显著扩大了数据的流通范围。

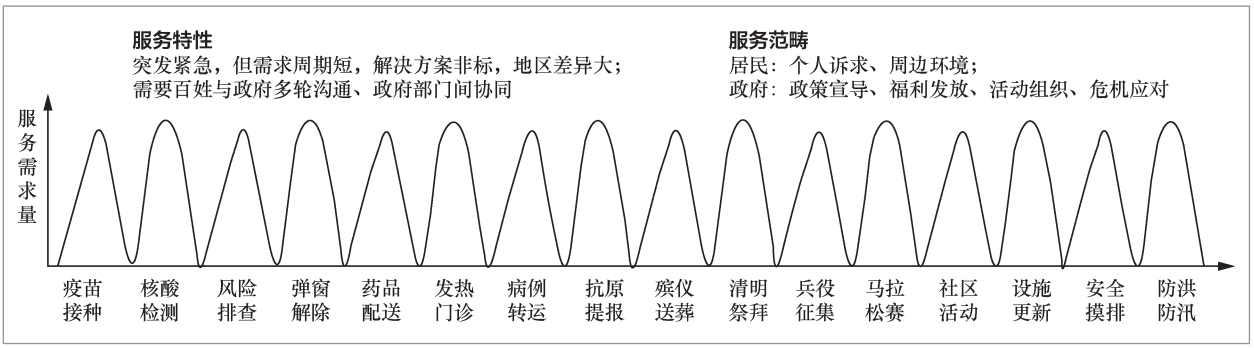

图 6 基层治理业务的特性和示例

6 结束语

本文设定了公共数据要素化的三大目标，阐述了构建公共数据要素的价值，介绍了该领域的研究进展，并分析了现有方法的不足和行业面临的挑战。针对以上问题，提出了人机智能协同的总体思路以及基于城市知识体系的数据要素构建理论，研发了承载该理论的数字化控件工具，并设计了智能术语学习和控件标题推荐算法。本文方法实现了数据与应用分离、数据要素的自动产生和更新、不同数据要素的自动连接三大目标，将数据要素的产生从方式事后、集中式的繁重治理方式转变为前置、分布式的自动产生方式，这有助于释放公共数据要素的价值，扩大数据要素的使用规模，促进我国数字经济发展。

参考文献:

[1] 朱扬勇. 依照数据用途界定公共数据[J]. 大数据, 2024, 10(3): 163–167.
ZHU Y Y. On public data[J]. Big Data Research, 2024, 10(3): 163–167.

[2] 郑宇. 城市知识体系[J]. 武汉大学学报(信息科学版), 2023, 48(1): 1–16.
ZHENG Y. The knowledge system for intelligent cities[J]. Geomatics and Information Science of Wuhan University, 2023, 48(1): 1–16.

[3] 中国信息通信研究院, 数据要素白皮书[R]. 2022.
China Academy of Information and Communications Technology. White paper on data elements[R]. 2022.

[4] 梅宏, 杜小勇, 金海, 等. 大数据技术前瞻[J]. 大数据, 2023, 9(1): 1–20.
MEI H, DU X Y, JIN H, et al. Big data technologies forward-looking[J]. Big Data Research, 2023, 9(1): 1–20.

[5] 郑宇. 城市治理一网统管[J]. 武汉大学学报(信息科学版), 2022, 47(1): 19–25.
ZHENG Y. Unified urban governance models[J]. Geomatics and Information Science of Wuhan University, 2022, 47(1): 19–25.

[6] 郑宇. 城市治理一网统管[M]. 北京: 机械工业出版社, 2022.
ZHENG Y. Unified management of urban governance network[M]. Beijing: China Machine Press, 2022.

[7] 郑宇. 政民互通：构建政府和居民之间的双向信息通道[J]. 大数据, 2024, 10(1): 127–140.
ZHENG Y. Building bidirectional digital channels between governments and citizens[J]. Big Data Research, 2024, 10(1): 127–140.

[8] 郑宇. 城市计算概述[J]. 武汉大学学报(信息科学版), 2015, 40(1): 1–13.
ZHENG Y. Introduction to urban computing[J]. Geomatics and Information Science of Wuhan University, 2015, 40(1): 1–13.

[9] 杜小勇. 大数据管理[M]. 北京: 高等教育出版社, 2019.
DU X Y. Big data management[M]. Beijing: Higher Education Press, 2019.

[10] 杜小勇, 陈跃国, 范举, 等. 数据整理: 大数据治理的关键技术[J]. 大数据, 2019, 5(3): 13–22.
DU X Y, CHEN Y G, FAN J, et al. Data wrangling: a key technique of data governance[J]. Big Data Research, 2019, 5(3): 13–22.

[11] 梅宏. 数据治理之法[M]. 北京: 中国人民大学出版社, 2022.
MEI H. Methods of data governance[M]. Beijing: China Renmin University Press, 2022.

[12] 王建冬, 于施洋, 窦悦. 东数西算：我国数据跨域流通的总体框架和实施路径研究[J]. 电子政务, 2020(3): 13–21.
WANG J D, YU S Y, DOU Y. East digital computing and west computing: research on the overall framework and implementation path of cross-domain data circulation in China[J]. E-Government, 2020(3): 13–21.

[13] LIU Y, LIU Y, LIU Z, et al. Federated forest[J]. IEEE Transactions on Big Data, 2020, 8(3): 843–854.

[14] 郑宇. 智能城市操作系统[J]. 中国计算机学会

通迅, 2020: 39–44.
ZHENG Y. Smart city operating system[J]. Communications of the CCF. 2020: 39–44.
[15] ZHENG Y. Urban computing[M]. Cambridge: MIT Press, 2019.
[16] 魏诚. 电子政务数据共享交换系统的设计与实现[D]. 南京: 东南大学, 2015.
WEI C. The design and application of e–government data sharing and exchange system[D]. Nanjing: Southeast University, 2015.
[17] ABRAHAM R, SCHNEIDER J, VOM BROCKE J. Data governance: a conceptual framework, structured review, and research agenda[J]. International Journal of Information Management, 2019, 49: 424–438.
[18] 梅宏. 数据治理之论[M]. 北京: 中国人民大学出版社, 2022.
MEI H. Theory of data governance[M]. Beijing: Chinese University Press, 2022.
[19] 杨孟辉, 杜小勇. 政府大数据治理: 政府管理的新形态[J]. 大数据, 2020, 6(2): 3–18.
YANG M H, DU X Y. Big data governance in governments: a new form of the government administration[J]. Big Data Research, 2020, 6(2): 3–18.
[20] 陆志鹏, 国丽, 乔亲旺, 等. 数据元件的结构要求: T/CIITA 406–2022[S]. 2022.
LU Z P, GUO L, QIAO Q W, et al. Requirement for structure of data components: T/CIITA 406–2022[S]. 2022.
[21] 陆志鹏, 国丽, 乔亲旺, 等. 数据元件安全审核要求: T/CIITA 506–2022[S]. 2022.
LU Z P, GUO L, QIAO Q W, et al. Audit requirement of data component for security: T/CIITA 506–2022[S]. 2022.
[22] 梅宏, 王亚沙, 赵俊峰, 等, 智慧城市 领域知识模型 核心概念模型: GB/T 36332–2018[S]. 2018.
MEI H, WANG Y S, ZHAO J F, et al. Smart city–domain knowledge model–core conceptual model: GB/T 36332–2018[S]. 2018.
[23] 史勇明, 张海梅, 张红卫, 等. 智慧城市 数据融合 第5部分: 市政基础设施数据元素: GB/T 36325–2019[S]. 2020.
SHI Y M, ZHANG H M, ZHANG H W, et al. Smart city–data fusion–part 5: Data elements of basic municipal facilities: GB/T 36325–2019[S]. 2020.
[24] 万碧玉, 吴丽丽, 马蓉, 等. 智慧城市 数据融合 第3部分: 数据采集规范GB/T 36625.3–2021[S]. 2021.
WAN B Y, WU L L, MA R, et al. Smart city–data fusion–part 3: Data acquisition specifications: GB/T 36625.3–2021[S]. 2021.
[25] 范举, 陈跃国, 杜小勇. 人在回路的数据准备技术研究进展[J]. 大数据, 2019, 5(6): 1–16.
FAN J, CHEN Y G, DU X Y. Progress on human–in–the–loop data preparation[J]. Big Data Research, 2019, 5(6): 1–16.
[26] HULSEBOS M, HU K, BAKKER M, et al. Sherlock: a deep learning approach to semantic data type detection[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2019.
[27] ZHANG D, SUHARA Y, LI J F, et al. Sato: contextual semantic type detection in tables[EB]. arXiv preprint, 2019, arXiv: 1911.06311.
[28] PAPOTTI P. Technical perspective of TURL[J]. ACM SIGMOD Record, 2022, 51(1): 32.
[29] 高珊, 袁宛竹, 卢卫, 等. 面向列语义识别的共现属性交互模型构建与优化[J]. 软件学报, 2023, 34(3): 1010–1026.
GAO S, YUAN W Z, LU W, et al. Construction and optimization of Co–occurrence–attribute–interaction model for column semantic recognition[J]. Journal of Software, 2023, 34(3): 1010–1026.
[30] NASAR Z, JAFFRY S W, MALIK M K. Named entity recognition and relation extraction: state–of–the–art[J]. ACM Computing Surveys, 2021, 54(1): 20.
[31] 臧根林, 王亚强, 吴庆蓉, 等. 智慧城市知识图谱模型与本体构建方法[J]. 大数据, 2020, 6(2): 96–106.
ZANG G L, WANG Y Q, WU Q R, et al. Model and construction method of the ontology of knowledge graph of smart city[J]. Big Data Research, 2020, 6(2): 96–106.
[32] 马亚中, 张聪聪, 徐大鹏, 等. 城市大脑知识图谱构建及应用研究[J]. 中文信息学报,

2022, 36(4): 48-56.
MA Y Z, ZHANG C C, XU D P, et al. Construction and application of city brain knowledge graph[J]. Journal of Chinese Information Processing, 2022, 36(4): 48-56.
[33] 杜小勇, 李彤, 卢卫, 等. 跨域数据管理[J]. 计算机科学, 2024, 51(1): 4-12.
DU X Y, LI T, LU W, et al. Cross-domain data management[J]. Computer Science, 2024, 51(1): 4-12.
[34] 黄丽华, 杜万里, 吴蔽余. 基于数据要素流通价值链的数据产权结构性分置[J]. 大数据, 2023, 9(2): 5-15.
HUANG L H, DU W L, WU B Y. Structural separation of data property rights based on data factor circulation value chain[J]. Big Data Research, 2023, 9(2): 5-15.

作者简介

郑宇（1979- ），男，博士，京东集团副总裁、京东智能城市研究院院长、京东科技首席数据科学家，IEEE Fellow，美国计算机学会杰出科学家，上海交通大学讲座教授，南京大学、香港科技大学等多所高校客座教授。先后担任人工智能顶尖国际期刊*ACM TIST*的主编、国家重点研发计划项目首席科学家及总负责人，以及ICDE及CIKM等多个国际会议的程序委员会主席。

易修文（1991- ），男，博士，京东城市数据科学家，入选2021年度北京市科技新星计划，主要研究方向为城市大数据智能。

齐德康（1993- ），男，西南交通大学计算机与人工智能学院博士生，主要研究方向为城市计算。

潘哲逸（1992- ），男，博士，京东智能城市研究院研究员，在国际顶级期刊或会议上发表论文10余篇，主要研究方向为城市计算、时空数据挖掘、深度学习。

收稿日期： 2024-04-25
基金项目： 国家自然科学基金项目（No.62076191）；北京市科技计划（No.Z211100004121008）
Foundation Items: The National Natural Science Foundation of China (No.62076191), Beijing Science and Technology Plan (No.Z211100004121008)