城市轨道交通大数据资源目录构建探究

潘　莹[1]　徐文洁[3]　颜彦文[2]　赵时旻[3]
（1. 万达信息股份有限公司，201103，上海；2. 上海久誉软件系统有限公司，201103，上海；
3. 上海申通地铁集团有限公司，201103，上海 //第一作者，工程师）

摘　要　针对目前城市轨道交通大数据资源目录构建中存在的分类不统一、描述不全面、服务不完善等问题，从资源分类及编码定义、核心元数据及资源目录服务模式等方面进行了探讨，提出了大数据资源目录构建方法及解决思路。

关键词　城市轨道交通；信息系统；大数据；资源目录
中图分类号　F530.7；TP274
DOI：10.16037/j.1007-869x.2021.06.031

Research on Construction of Big Data Resource Directory of Urban Rail Transit

PAN Ying, XU Wenjie, YAN Yanwen, ZHAO Shimin

Abstract　Targeting the problems in the construction of big data resource catalogue of urban rail transit such as ununified classification, incomplete description and unsatisfying service, discussions are carried out on topics of resource directory and coding definition, core metadata and service mode of resource directory in urban rail transit, and solutions to support the construction of big data resource directory are put forward.

Key words　urban rail transit; information system; big data; resource directory

First-author's address　Wonders Information Co., Ltd., 201103, Shanghai, China

城市轨道交通数据是交通信息资源的重要组成，蕴含了大量有价值的信息。然而，城市轨道交通经过多年的发展，其数据内容、形式复杂多样，数据资源的结构划分、资源描述以及资源目录服务缺乏统一的行业标准，极大地影响了其数据资源的价值利用。因此，建立一套有效、合理的城市轨道交通大数据资源目录，能够促进数据资源的有效组织和准确描述，帮助打破城市轨道交通信息系统建设中出现的信息孤岛现象，促进城市交通大数据的业务融合，提升智慧交通的服务水平。

关于数据资源目录构建，我国交通行业主管部门已颁布过一系列行业标准和指导意见。2017 年交通运输部办公厅发布了《交通运输政务信息资源目录编制指南（试行）》[1]（以下简为《目录编制指南》），对资源目录编制进行指导；2020 年交通运输部发布行业标准 JT/T 747.3—2020《交通运输信息资源目录体系 第 3 部分：核心元数据》[2]（以下简为《核心元数据》），规定了核心元数据的描述方法、数据内容、扩展要求及值域代码。上述标准从元数据描述、资源分类方法以及共享机制等层面为资源目录构建提供借鉴和指导，但目前针对大数据资源以及城市轨道交通细分行业尚无详细的标准规范。因此，需深入探讨城市轨道交通大数据资源目录构建所面临的具体问题，研究构建方案，并提出解决思路。

1　大数据资源目录构建的难点

城市轨道交通大数据具有数据量大、种类繁多、各业务口径定义复杂等特征。因此，大数据资源目录构建过程中存在以下问题和难点。

1.1　大数据资源分类不统一

城市轨道交通大数据的内容复杂，应用场景丰富，不同资源目录构建人员往往仅从各自应用的角度对数据进行目录划分和定义。表 1 为城市轨道交通大数据资源的分类。

表 1 中的数据状态、数据来源、数据格式、数据安全、业务对象、业务领域、管理目标及数据服务等维度对于不同的构建人员有着不同的分类和定义侧重。业务人员更关注于快速获取业务领域数据，信息人员关注于数据质量安全分析数据，而管理人员则关注管理目标分析预警数据。不同的关注侧重导致了数据分类内容的差异性，也造成了同一资源在不同分类中交叉覆盖、编码规范各异等问题。资源目录分类难以统一，造成用户的资源视图不清晰，影响用户查询和检索资源的效率，降低大数据资源目录共享水平。

表 1　城市轨道交通大数据资源的分类

划分维度	类型	具体内容
数据状态	静态数据	地图数据、线网站点、设备等数据
	动态数据	运营数据、设备监测数据、网络安全数据、企业管理数据、公共服务数据等
数据来源	行业产生数据	规划、建设、运营、维护、企业管理、技术研发、投融资
	关联行业数据	城市规划、公共交通、公安、气象、环境、地质等数据
	公众及外部数据	信访、舆情、供应商、造价等数据
数据格式	结构化	设备管理系统、项目管理系统、运营生产系统产生的数据
	半结构化	运营日志等数据
	非结构化	地铁预案、技术标准、视频等数据
数据安全	公开数据	面向公众及企业无条件开放数据，如研究性样本数据
	不公开数据	暂时不予开放的数据
	授权公开数据	按照一定的授权程度向不同用户开放的数据
业务对象	按对象模型划分	规划、建设、运营、维护、企业管理、技术研发、投融资等
业务领域	按业务类型划分	围绕基础业务对象如线路、车站、乘客、设备等数据
管理目标	按管理目标划分	围绕服务、效益、安全等管理数据
数据服务	按服务方式划分	围绕存储类数据服务、分析类数据服务、可操作类数据服务
更多分类	按更多需求划分	不同用户按照各自维度进行划分

1.2　大数据资源描述不全面

大数据资源特征的描述是资源共享和交互的基础。在大数据时代，借助于元数据了解数据元素含义和上下文的需求越来越强烈。当前国际通用元数据标准主要有美国国家信息标准协会（NISO）的都柏林核心元素集和 W3C（万维网联盟）发布的 DCAT（数据目录词汇表）正式推荐标准。《目录编制指南》提出，核心元数据包括必选项、可选项及扩展项等 3 部分。其中，必选项包括信息资源分类、信息资源名称、信息资源代码、信息资源提供方、信息资源提供方代码、来源系统、信息资源摘要、信息资源格式、信息项信息、共享属性、共享方式、开放属性、更新周期及发布日期；可选项主要包括来源数据库、信息资源格式、信息项信息、开放属性、关联资源代码及数据元编号；扩展项指根据目录编制单位的实际情况和需要添加的元数据项。《目录编制指南》聚焦交通运输政务信息资源描述普适性、通用性标准，但无法详细全面地描述城市轨道交通大数据资源及行业特征，因此，造成了城市轨道交通行业的数据资源无法实现更好地共享和交互。

1.3　大数据资源目录服务不完善

城市轨道交通大数据服务范围不仅仅包括企业用户、行业用户，还涉及公共服务用户以及大数据研究者等专业用户。服务的内容、模式主要归纳为 4 个层面：

1）企业服务层。企业服务层主要面向企业内部用户。企业内部大数据应用将依托核心业务领域开展，如运营评估与应急、客流分析及预测、资产设备状态与监控、乘客行为分析以及线路规划等方面。企业通过大数据分析和处理技术，挖掘和使用数据资源，精准掌握业务状态、发展规律及趋势，形成大数据驱动的业务创新模式，服务于安全、效率、服务等各项关键绩效指标。

2）行业服务层。行业服务层主要面向政府及行业主管部门、联动单位。主管部门关注行业创新、地域规划开发的统筹协调、民生关怀等内容。政府及行业主管单位对城市轨道交通规划、建设、运营等大数据进行分析和研究，指导行业的健康发展。联动单位需要及时共享天气、客流、舆情、联动任务等公共信息，提升城市整体协作水平，助力智慧城市建设。

3）公共服务层。公众服务层主要面向乘客和供应商。乘客需要在出行场景中获取持续的大数据服务，如线路推荐、候车预测、LBS（基于位置的服务）、出行建议及安全提示等。供应商需要获取或定制招投标信息、实时动态资讯服务。

4）专业服务层。专业服务层主要面向专业的大数据研究机构或者人员。城市轨道交通大数据蕴含极大的价值，需要该产业链上下游单位及专业研究者的协作开发。大数据研究机构或人员往往关注如何获取样本数据、开放算法或者可共享的分析成果及案例等。

综上所述，大数据背景下城市轨道交通大数据资源服务的范围、内容及模式发生了巨大改变。大数据资源目录服务体系需进一步完善和深化，从而为用户提供更加便捷、安全和个性化的服务。

2　大数据资源目录构建探讨

为解决城市轨道交通大数据资源目录构建工作所面临的一系列难点，本文在借鉴相关标准的基础上，从资源分类及编码定义、核心元数据定义及资源目录服务定义等方面进行探讨。

2.1　资源分类及编码定义

信息资源分类的方法一般采用混合分类法。如《目录编制指南》采用混合分类法时，以信息资源涉及的行业管理及其业务范围作为两个基本分类依据，并在业务范围内从管理对象、行为主题和信息类别等 3 个不同维度进行信息分类[1]。混合分类有利于数据资源按不同维度进行组织，从而提升大数据资源的可用性，满足不同用户获得相应资源以及应用不同场景的需求。

借鉴混合分类框架，结合城市轨道交通行业大数据资源特征对数据进行分类。随着城市轨道交通的发展，管理者对地铁运营安全、运维效率和服务质量越来越重视[3]，因此，管理决策者需对资源进行跨行业整合，以便从大数据资源中挖掘有价值的信息，赋能企业发展。城市轨道交通大数据资源，不仅汇聚融合行业及关联数据本身，还包括支撑大数据常用工具、算法以及分析成果和知识供不同的用户使用。该类资源无法简单地用现有信息类别中的“统计信息”来描述。因此，大数据资源目录分类需在《目录编制指南》行业管理及业务范围两个基本分类依据上进行完善，通过增加大数据资源服务类维度来扩展原有信息分类的范围（见图 1）。

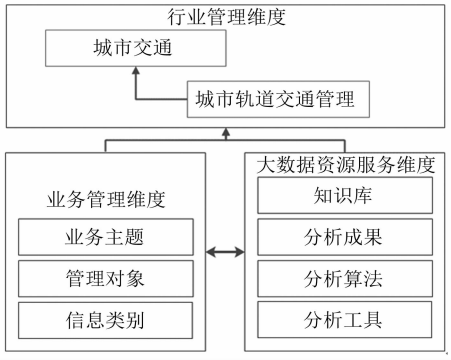

图 1　城市轨道大数据资源分类维度

如图 1 所示，行业管理维度中城市轨道交通大数据资源属于城市交通行业中的城市轨道交通管理分类；业务管理维度中建议增加综合监管分类，用于描述管理者关注的企业综合运行信息；而大数据资源服务维度则是对大数据的各类成果工具、算法、成果及知识等进行标识，从而促进大数据向业务驱动转化。基于此分类思路，本文设计的城市轨道交通大数据资源可按照类-项-目-细目进行细化（见图 2）。

图 2　城市轨道大数据资源分类模型

城市轨道交通行业资源分类编码是数据资源共享的基础,采用现有交通行业的资源分类标准编码有利于提升城市轨道交通大数据资源的辨识度,从而解决跨行业高效共享的问题,因此,行业分类、管理对象、业务主题的编码采用《目录编制指南》中规定的标准编码,大数据服务资源类采用自定义编码。为了解决分类编码不同但资源相同的问题,在元数据描述中增加"关联资源代码",建立不同编码间的关联链接。

2.2 核心元数据定义

为了使大数据资源的描述更加全面准确,建议从现有的《核心元数据》标准及扩展原则出发,围绕城市轨道交通大数据关键特征,以能准确而全面地描述城市轨道交通大数据资源为目标,来弥补核心元数据中大数据及城市轨道交通特征描述的不足。

在大数据特征上,城市轨道交通大数据具备典型的 4V 特点:①数量巨大(Volume),如客流、列车运营等动态实时类数据数量巨大;②数据复杂多样(Variety),如数据格式多样、结构不一、存储分散;③对数据实时性(Velocity)要求较高,如设备异常状态需及时反馈至指挥人员;④城市轨道交通的价值密度低但价值高(Value),决策人员需要从海量数据中获取分析或进行预测。

随着大数据的广泛应用,数据的拥有者和管理者分离,其生命周期变为产生、传输、存储及使用[4],因此,城市轨道交通大数据在质量、安全、隐私及服务等方面的描述需求变得日益突出。在原有核心元数据描述中的"来源系统"中定义了数据源的定义,大数据的资源有可能是多源系统汇聚后的成果,所以需要标识是否为多源数据、数据来源的标签等。针对数据的质量,大数据资源目录描述一般为清洗后的数据,对于数据质量本身的描述较少,资源使用人员对获取数据的质量无从了解,影响数据使用的效果,建议增加精确性、完整性、有效性及清洗的程度等类别来描述数据的质量;针对隐私安全信息,《核心元数据》通过"共享属性"、"共享方式"、"开放属性"等进行描述,但是上述信息还不够完善,建议增加隐私的条例、加密等具体描述,确保数据资源的安全性描述更加精细。在数据服务属性方面,随着未来大数据服务更加深入和广泛的应用,该项服务将被不确定的用户进行封装和调用,数据服务的属性需要确保唯一的标志,同时其服务内容、服务许可、服务质量等信息需进一步描述。

在行业特征上,城市轨道交通大数据具备复杂的时空属性、动态性、周期性等特征[5]。例如,若空间参照系不同,地铁车站、线网、列车动态运营等数据资源将无法有效叠加应用,因此,可以沿用《核心元数据》中的"时间范围信息"并扩展"坐标系信息"等空间属性对城市轨道交通基础空间对象进行描述。属性对城市轨道交通基础空间对象进行描述。此外,城市轨道交通大数据和外部多源异构数据相关性较大,如地铁客流预测服务需要考虑和天气、节假日、时段、站点位置、外部重大活动等外部多源异构数据的相关性。与客流预测相关的大数据资源服务进行描述时需要描述关联主题,以便资源使用者可以清晰地了解和预测模型的运行影响参数。因此,本文融合大数据和城市轨道交通行业数据资源特征对核心元数据进行扩展,确保用户在利用大数据资源时,能够清晰地了解数据资源的细节(见表 2)。

表 2 城市轨道大数据核心元数据扩展属性

扩展项	关键元素	说明
数据源属性	多源数据标识	是否来源于多数据源
	数据源的标签	多数据源的标签
数据质量属性	精确性	数据源中是否有部分数据错误
	完整性	数据是否存在缺失
	有效性	如法规、标准等是否有效/废除
	冗余性	数据是否存在重复纪录的情况
	清洗程度	数据被清洗的情况
数据隐私属性	隐私条例	与该数据源的隐私限制相关的规定
	是否加密	数据源的数据是否加密及加密方法
数据服务属性	唯一标志	唯一标识该数据服务
	服务描述	对服务内容的具体描述
	服务质量	性能及可靠性
	服务类别	查询型、分析型、可视化
	服务许可	服务许可信息及加密信息的描述
	关联服务	关联的服务资源

2.3 资源目录服务模式定义

大数据背景下,数据服务模式已经发生了改变,原始数据查看和下载以及基础性服务接口,已经无法满足未来大数据的应用场景,因此,需要定义一套有效、合理的资源目录服务以支撑大数据服务模式。对于资源目录服务,一方面,各类用户需要定制化、专业化及方便灵活的数据资源服务;另一方面,管理者需要应对大数据带来的安全、隐私等问题带来的挑战,制定管理制度确保资源目录体

系持续、有效、规范地运行。

大数据服务是一种数据使用模式,是在对大数据统一建模的基础上,将各类数据操作进行封装,对外提供无所不在的、标准化的、随需的检索、分析或者可视化的服务交付。大数据服务不仅仅是一种新技术,也是一种新的数据资源使用模式和服务经济模式[6]。大数据资源目录服务建设借鉴了大数据服务理念,首先完成城市轨道交通大数据资源目录树的构建,然后结合业务需求,将数据资源组合封装成个性化服务,如主数据服务、基础报表服务、风险预警服务、关键绩效指标服务、开放性研究样本服务及共享算法服务等。

用户可以依据关键字来检索数据目录资源,也可以按照不同的管理对象、业务主题、资源服务方式等维度浏览、查看、下载资源。大数据资源广泛且数据结构复杂,为保证对超大量索引数据的快速搜索支持,本文设计分布式的存储方式对元数据的目录进行部署,采用索引文件分块技术,并支持批量索引和复合搜索。非结构化文件的检索设计有两种方式,一种是进行全文快速检索,支持用户使用布尔逻辑运算、部分匹配、通配符、输入内容自动补全等功能进行模糊查询,分析文本文件内的具体内容,并且支持在查询的结果中进一步分析筛选。另一种方式是高级搜索,即通过非结构化数据资源的属性对非结构化文件进行搜索服务。该搜索方式包括基本搜索、文件夹和元数据搜索、混合搜索等,搜索条件丰富,同时能够满足主要搜索需要。

为了确保搜索的安全性,对大数据资源的安全级别与系统设置,用户依据权限通过对大数据资源进行访问、调阅、申请、利用等操作进行鉴权管理。在非授权的情况下,数据资源将不会被搜索到或者无法预览细节,搜索的范围与结果会被系统安全权限控制,保障了大数据的出口安全。

3　结语

本文结合上海申通地铁集团有限公司的大数据项目建设,分析了与大数据资源目录构建密切相关的大数据内容、特征及应用等问题,深入探讨了城市轨道交通大数据资源目录构建中的资源分类及编码定义、核心元数据定义、资源目录服务模式定义等核心问题。本文提出的大数据资源目录构建方法及思路,可为上海申通地铁集团有限公司的大数据中心数据规划提供基础参照,对城市轨道交通大数据资产管理及未来经营模式具有借鉴意义。

参考文献

[1] 国家发展改革委,中国网信办. 交通运输政务信息资源目录编制指南(试行)[EB/OL]. (2017-07-13)[2019-04-22]. http://www.ndrc.gov.cn/zcfb/zcfbtz/201707/W020170713603384554898.pdf.

[2] 中华人民共和国交通运输部. 交通运输信息资源目录体系 第3部分:核心元数据:JT/T 747.3—2020[S]. 北京:人民交通出版社,2020:3.

[3] 世界级超大规模地铁网络运营背后的"管理哲学"[J]. 世界轨道交通,2016(8): 16.

[4] 张尼. 大数据安全技术与应用[M]. 北京:人民邮电出版社,2014.

[5] 李得伟,张天宇,周玮腾,等. 轨道交通大数据运用现状及发展趋势研究[J]. 都市快轨交通,2016(6): 1.

[6] 韩晶. 大数据服务若干关键技术研究[D]. 北京:北京邮电大学,2013: 70-96.

(收稿日期:2019-08-22)

(上接第 140 页)

输、安装过程的荷载、吊装基础等预留。

参考文献

[1] 中国人民共和国住房和城乡建设部,中华人民共和国国家质量监督检验检疫总局. 地铁设计规范:GB 50517—2013[S]. 北京:中国建筑工业出版社,2013: 258.

[2] 中交城市轨道交通设计研究院有限公司,中铁二院工程集团有限责任公司. 湖涌停车场初步设计修改报告[R]. 佛山:中交城市轨道交通设计研究院有限公司,2017.

[3] 金永乐,张子健. 改善上盖地铁车辆段运营条件设计新思路[J]. 都市快轨交通,2018(3): 6.

[4] 刘仁猛,陈苏,皇甫学斌. 城轨车辆段及上盖物业给水及消防工程设计[J]. 给水排水,2013(11): 69.

(收稿日期:2019-09-01)