政府数据开放注册管理元数据研究

赵盼云 耿骞 柏林
（北京师范大学政府管理学院，北京 100875）

摘要：为解决政府开放数据集界定不清晰，数据集更新不及时且缺乏有效的政府部门间、政府部门和社会组织间数据交换机制问题，本文对政府数据开放注册管理进行研究，探讨注册管理中相关要素和注册过程，并在此基础上，对与数据注册关系密切的元数据标准进行探讨。通过对注册管理过程中数据集的描述需求分析，剖析国外政府数据的开放标准DCAT、POD和DCAT-AP，提出针对我国政府数据开放注册管理元数据标准建设的建议。

关键词：政府数据开放；注册管理；元数据标准
中图分类号：G254
DOI：10.3772/j.issn.1673-2286.2018.05.008

1 政府开放数据的研究背景与研究现状

1.1 研究背景与问题

政府数据指所有产生于政府内部或外部，对政府活动、公共事务和公众生活有影响、有意义的数据资源统称[1]。政府在进行国家和地方管理活动时生成大量数据，这些数据有巨大的开发潜力和应用价值。政府数据的有效管理和利用对提高政府工作效率、促进社会经济发展和产业升级具有重要意义。

政府数据开放为大数据产业发展带来契机，对于数据的交换、利用和创新具有重要意义。2009年以来，美国和英国相继开展政府数据开放运动，发布政府数据开放平台。随后，加拿大、法国、挪威、肯尼亚、韩国和新加坡等国家也建立政府数据开放网站，开启全球政府数据开放的浪潮，为政府数据共享和交换提供极大便利。上海市于2012年6月首先上线试运行“上海市政府数据服务网”，随后北京市、武汉市等也陆续推出开放门户。

尽管我国自2011年开始推动政府数据开放建设工作，但从近年发展情况看，进展不甚理想，主要体现为数据集数量少、内容易变的数据集更新不及时；同时，与此相关的大数据交易平台也存在数据管理及产权等方面的问题[2]，给相关数据交易带来障碍。政府数据开放工作推进一个主要问题是数据开放门户中相关的管理功能和机制不完善，平台交互性不强，影响政府各部门、社会组织和公众通过平台进行数据发布、描述、发现和使用。该问题来源于对数据开放管理流程和机制的认识与设计，深层原因在于采用何种方法有效进行数据的规范，如何明晰数据提供者与使用者的利益与责任，从而提高各参与方的积极性。

解决以上问题，既需要数据开放相关制度与机制的完善，也需要数据管理方法与技术的提高。从数据管理方法上看，数据集的注册管理是基础的、能较好带动相关问题解决的工作。注册是用于对权威的、集中控制的信息进行存储的一种管理机制[3]。注册管理的一项基础和重要工作是制定相关的元数据标准。本文从政府开放数据注册管理入手，对注册管理中的参与者和功能需求进行分析，并通过对主流政府数据开放元数据推荐标准DCAT（Data Catalog Vocabulary）分析，结合国外相关研究和实践，借鉴美国、欧盟对DCAT扩展和改造的经验，为制定我国政府数据开放注册管理元数据标准提供建议。

1.2 研究与实践现状

数据注册管理与相关元数据标准是来源于实践领域的具体问题，目前，美国、英国及欧盟等国家和组织都已开始相关方面的实践工作。这些国家和组织或

将政府数据注册管理集成到政府数据开放平台中，或使用专门的数据注册管理平台完成相关数据集的注册工作。在元数据标准方面，目前较普遍使用的是W3C政府关联数据工作组于2014年1月发布的DCAT[4]，美国、英国及欧盟等国家和组织或直接采用该标准，或在该标准基础上进行修订完善，形成该国家或组织的标准[5]。如美国基于DCAT制定了POD（Project Open Data Metadata Schema），目前已更新到v1.1版[6]；欧盟以DCAT为基础，从提高欧盟电子政务系统的语义互操作性出发，通过其“公共部门互操作性方案”行动计划1.1编制了DCAT-AP（Data Catalog Vocabulary-Application Profile）[7]，目前DCAT-AP已经在超过15个政府数据开放平台中使用[8]（如挪威、荷兰等国家的政府数据开放平台）。在系统平台方面，目前政府开放数据注册管理系统使用最广泛的是CKAN（Comprehensive Knowledge Archive Network），该系统是由国际开放知识组织研发的基于网络的开放源代码管理系统，用于存储和分发开放数据，政府和个人用户都可使用，同时为官方和社区数据门户网站提供支持，目前很多国家和组织都使用该系统。

政府数据开放及相关的数据注册、元数据标准更多属于实践活动范畴，目前该方面的理论研究较有限，更多的研究集中在元数据相关标准方面。其中，如Tygel等[9]研究政府数据开放的元数据解决方案，认为跨门户元数据不能只用数据目录词汇表解决，因此，开发实施了一种开放数据门户中的标签协调方法，包括与个人门户相关的本地操作，以及在单个门户网站上添加语义元数据层的全局操作，旨在提高单个门户中的标签质量；Neumaier等[10]对260个开放数据门户进行质量指标分析；Martin等[11]采用Berners-Lee's的关联数据五星标准分析PublicData.eu数据集目录。此外，也有一些研究通过对元数据模型进行调整以增强其性能，如Brümmer等[12]基于DCAT和VoID词汇表的数据模型创建新的DataID，以进一步提高RDF数据集的可发现性；Assaf等[13]在对CKAN、DKAT、Public Open Data、Socrata、VoID、DCAT和Schema.org7个元数据调查的基础上，提出协调数据集模型。

国内对政府数据开放研究主要集中在对国外政府数据开放元数据的调研、介绍及资源描述方面。其中，武琳等[14]梳理美国等国家和欧盟的相关元数据政策和标准，对元数据格式、元数据框架、元素、数据目录词表、受控词表等方面进行比较分析，认为设计元数据标准时需关注3个要素，即数据目录词表的支持、受控词表的有效使用、面向元数据的关联本体；赵蓉英等[15]调查data.gov.uk的两种类型元数据，即面向网站数据资源的CKAN格式记录和GEMINI地理空间元数据标准；翟军等[16]对国内外政府数据开放元数据方案进行介绍和研究。在政府数据开放中资源描述方面，也有研究在分析政府数据开放特点的描述要求基础上，引入DC、VoID和DCAT等元数据标准对数据资源描述进行研究[17-18]；黄如花等[19]通过对英国等国家和欧盟的政府数据开放门户及其相关公共部门的元数据描述规范进行调研后，建议我国开放政府数据描述规范采用国际通用标准拟定国家元数据标准草案、适应网络特点编写控制词表、统一元数据格式。

总体来看，我国政府数据开放建设起步较晚，相关理论研究还不够丰富。同时，在数据开放的组织和管理层面上，国外政府数据开放通常从国家层面启动和推进，而我国由于管理问题的复杂性和特殊性，通常从省或市一级进行试点。目前，在我国已开展的政府数据开放工作中，实施较好的有北京市、上海市、武汉市等。但这些地区的实践并没有开展相应的数据注册活动，也没有统一的元数据规范。在元数据管理和描述上，基本各自为政，且都是一些简单的元数据元素（如资源名称、摘要、机构名称等）。由此可以看出，我国对开放政府数据注册及管理元数据标准研究尚处于初始阶段，还没有对制定我国元数据标准进行深入的研究，也没有从注册管理角度对政府数据开放中的元数据标准进行相关研究，本文以此为出发点进行政府数据开放注册管理的研究，以期为后续研究提供借鉴。

2 政府数据开放注册管理

2.1 政府数据开放注册的参与者和过程

数据集注册的目的是为更好地进行大数据管理。注册需要平台来支持有关活动，数据集提供者和数据集使用者通过注册管理平台完成相关工作，三者的关系如图1所示。

其中，数据集提供者对数据集进行描述，并将元数据提交到注册管理平台的“数据集元数据目录”中。同时，数据集提供者需要及时更新并发布数据，使数据集使用者能够进行信息查询和下载数据；数据集使用者通过注册管理平台查找相关数据集，并根据查询

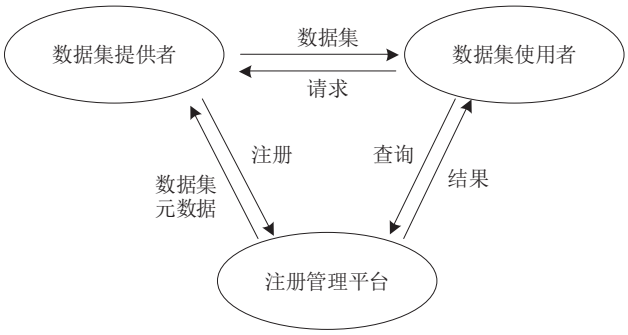

图1 注册活动的参与者及相关关系

结果所提供的相关信息获得数据集。注册管理平台是“数据集元数据目录”的载体，提供相关的注册管理功能。具体包括以下方面。

（1）用户注册与身份认证。注册管理平台需要对使用平台的用户进行身份认证，并进行用户管理。对于政府部门的个人或机构可以通过统一方式进行批量注册认证。

需要注意的是，并非所有用户都可以进行注册和托管数据集。如美国的Inventory.data.gov平台注册与OMB MAX平台登录集成在一起，而OMB MAX平台只有拥有联邦政府电子邮件的人员可以自行注册，其他人员则需要联系MAXSupport@max.gov进行注册。OMB MAX是为政府范围内的联邦机构和合作伙伴提供高级合作、信息共享和数据收集、发布、分析服务的平台[20]。因此，对于除政府部门外的数据集，需注意数据集来源的真实性和有效性，提高对这些数据集的注册门槛，可以在一定程度上保证数据集质量。

（2）数据集注册管理。政府数据开放活动中的数据来源于政府部门数据和政府部门外的相关数据。其中，前者是政府数据开放中的主要数据来源，后者是来自个人、其他社会组织的数据资源，通过政府数据注册管理平台登记并提供共享的数据。

对于产生于政府部门的数据集，其注册管理及数据获取包括三个步骤。首先，政府部门根据政府制定的开放数据政策提供所需数据集清单及元数据，数据集的元数据标准来自相关权威机构制定的元数据规范；其次，在部门机构网站上提供用户阅读和机器可读的公共数据元数据清单；最后，政府数据开放统一平台对机器可读元数据进行收割，并定期更新数据目录。该过程如图2所示。

（3）制定数据集接收政策。政府数据集须经过相关政府部门审查，而非政府部门人员提交的数据集信息无法统一控制，需要制定数据接收政策来限制数据内容，并考虑数据集的版权和密级等问题。接收政策应规定数据集合理合法的使用范围，并要求数据集的完整性。数据集内容不能含有违法信息，如涉及国家机密、影响他人权利的信息，还应在合适的开放许可协议下进行开放。

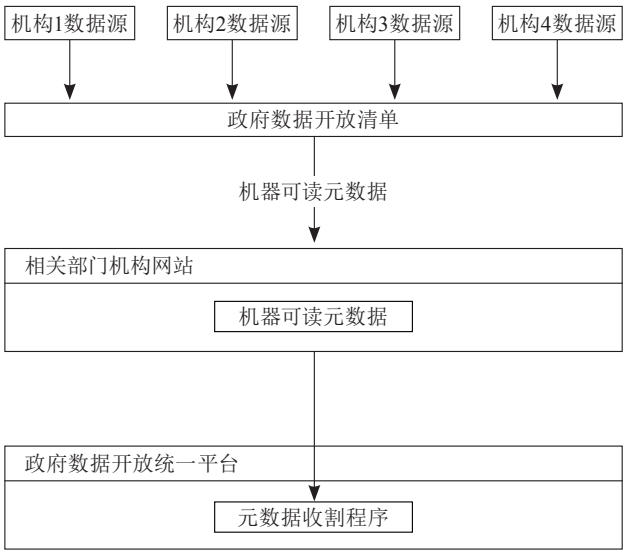

图2 政府数据开放数据集注册管理流程

（4）提供查询服务。查询服务包括目录式浏览查询和字段检索。注册管理平台应向数据集使用者提供尽可能多的查询方式和尽可能详细的数据集元数据相关信息。

（5）维护和管理本地元数据存储。验证已接收的数据集元数据是否符合模型和格式要求，如果不符合，将错误日志返给数据集提供者。同时根据数据集提供者提供的数据集清单，核对各机构“数据集元数据目录”提交进度。此外，还负责对数据集元数据的修改、删除和更新。

（6）提供软件工具。开放注册管理平台需要提供3种工具。①数据集元数据验证器。如以DCAT为模型的JSON格式数据集元数据，需要数据集元数据验证器验证其是否符合DCAT的要求。②数据集元数据格式转换器。若数据集元数据直接写成JSON格式比较复杂，可将数据集元数据转换成CSV格式，再通过数据集元数据转换器转为JSON格式。③数据集元数据测试收集器。用来测试将数据集元数据提交给数据集注册管理中心时可能出现的问题。

（7）数据集反馈。数据集使用者在使用数据集时发现的错误，可通过注册管理平台反馈给数据集提供者。

（8）数据集托管服务。对于没有数据集存储网站的个人或组织，数据注册管理平台也可以提供托管服务，具体的服务模式可根据情况而定。

此外，应创建另一个与政府数据管理平台基础架构相同的网站，用于协助没有政府数据开放建设方案的机构来创建和维护自身数据库。未进行政府数据开放的政府机构和非政府数据集提供者可通过注册将自身数据集托管给此网站，通过此网站可以增加、编辑、删除数据集。此网站还应支持数据集以政府数据注册管理平台元数据标准导出的元数据。

2.2 数据集描述需求

通过以上分析，可知数据集应主要对以下方面进行描述。

（1）对数据集的描述。除标题、简介、关键词、发布日期等信息的描述外，还需要4个信息。①数据集的开放许可协议。每个数据集都应是合理合法的，对于使用数据集是否需要注明出处、是否允许修改等做进一步说明，使用户对数据集的利用更加规范，保障数据集提供者的利益。②数据集密级。元数据描述时，应对数据集的密级进行说明，避免由于政府数据开放引起负面事件。③数据集内容所涵盖的时间和空间范围。④数据集主题。

（2）对数据集发布信息的描述。主要包括对数据集标题、发布日期、修改日期等方面的描述，用于记录数据集的发布状态。

（3）对发布者个人或机构信息的描述。需要对数据集提供者的名称、主页、联系方式等信息进行描述，方便数据集使用者对数据集提供者进一步了解。

（4）对数据集发布形式的描述。数据集可能有不同格式（如XML、RDF格式等），也可能只是一个网页链接，所以要对可用的发布格式、大小等信息进行描述。

3 政府数据开放元数据标准

开放政府数据包含大量来自政府不同部门，乃至政府外部的数据，这些数据与政府行政及公共事务密切相关，涵盖气象、水文、交通、健康、医疗等多个方面，且数据集的形态各异、结构和数据格式不同。同时，数据通常是动态的，需要不断补充和更新。为保证有效地进行数据集的注册、发现、收割和使用，需要制定有效的政府数据开放元数据标准，以对数据集进行准确的描述。目前，在国外政府数据开放实践中已经出现若干相关标准，这些标准对制定我国政府开放数据标准具有较强的借鉴意义。

3.1 DCAT

如前文所述，DCAT是W3C政府关联数据工作组制定的政府数据开放元数据推荐标准，其提供了一个数据描述模式，具有较完善的数据集描述结构和形式，其描述体系由类及其属性组成，如图3所示[4]。

DCAT词汇表涉及7个类，分别是dcat:Catalog，dcat:Dataset，dcat:Distribution，dcat:CatalogRecord，skos:ConceptScheme，skos:Concept，foaf:Agent。其中前4个为DCAT自定义类，分别是表示数据集目录的Catalog类，数据集的Dataset类，数据集发布形式（例如可下载文件、RSS或提供数据的Web服务）的Distribution类，描述数据集条目出处信息的CatalogRecord类。CatalogRecord是可选类，其属性主要复用自其他的词汇表。

DCAT复用一些其他词汇表，在使用时也可以引用这些词汇表外的词汇表，具体如表1所示。

DCAT引用表1中的元素，尤其是DCMI Metadata Terms，通过复用元素增加数据目录间的可操作性。使用DCAT描述数据目录中的数据集，应用程序能使用多个目录的元数据，并支持分布发布目录和跨站点的联合数据集搜索。

3.2 POD

POD是美国为进行数据交换而制定的元数据标准，目前版本是v1.1。POD通过字段描述相关信息，其作用相当于DCAT中的类。POD的字段呈现层次型字段、子字段的嵌套结构，如图4所示[6]。

POD v1.1的5个字段分别是Catalog、Dataset、ContactPoint、Distribution和Publisher。每个字段又有子字段（相当于属性），而Dataset字段下的contactPoint、publisher子字段也有其子字段。在Catalog字段中，除3个子字段外，还包括@context、@id和@type3个子字段，它们是JSON-LD的关键字。在其他字段中也存在类似情况。使用POD描述数据集时，最终以JSON格式进行描述。Publisher字段的右半部分

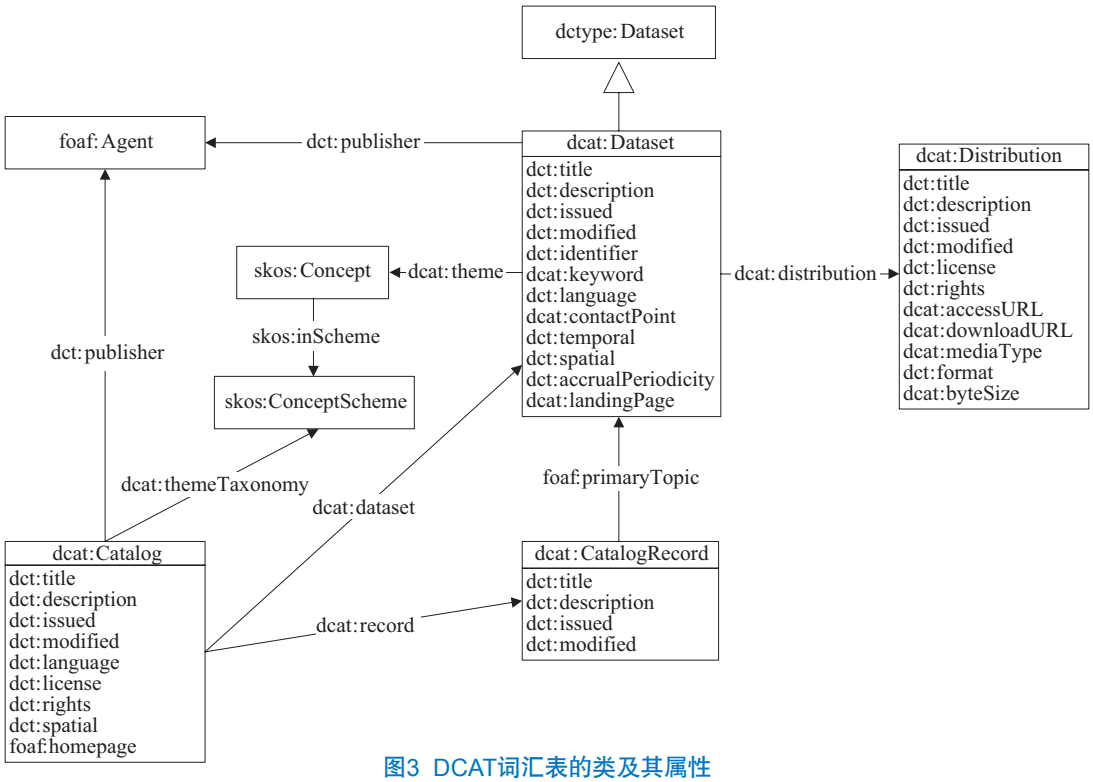

图3 DCAT词汇表的类及其属性

表1 DCAT复用的词汇表

前 缀	词汇表名称	词汇表主要内容
dct	DCMI Metadata Terms	由都柏林核心元数据倡议组织维护的元数据规范，包括属性、词汇编码方案、语法编码方案和类别
dctype	DCMI Type Vocabulary	提供通用的跨域术语列表，用于标识资源类型
foaf	Friend of a Friend Vocabulary	使用Web连接人员和信息的词汇表
rdf	The RDF Concepts Vocabulary	用于在Web中表示信息的框架
rdfs	The RDF Schema Vocabulary	RDFSchema是对RDF基础词汇的扩展，为RDF提供数据建模词汇表
skos	Simple Knowledge Organization System	通过语义网共享和链接知识组织系统的通用数据模型
vcard	vCard Ontology	vCard是用于描述人员和组织的规范
xsd	XML Schema Definition	W3C发布的推荐标准，指出如何形式化描述XML文档的元素

箭头指自身的嵌套层级关系，具体见图5。

这段代码显示了一个publisher实例的具体身份，是美国政府（U.S. Government）“General Services Administration”下属的“Office of Citizen Services and Innovative Technologies”的一个项目，即“Widget Modulation Program”。

POD的字段分为必备（required）、适用则必备（required if applicable）和扩展（expanded）3种类型，其中扩展字段是可选择的。

POD可映射到其他元数据标准，如DCAT、Schema.org、CKAN元数据[21]、ISO 19115和CSDGM [22]。其中，Schema.org词汇表由Microsoft、Yahoo!、Google和Yandex合作制定，目的是创建一种能使主要搜索引擎都支持的结构化数据标记架构，以帮助搜索引擎理解网页内容，并提供更丰富的搜索结果[23]。ISO 19115是地理信息元数据标准[24]，用于描述地理信息和相关服务的国际通用模式，提供有关数字地理数据的识别、范围、质量、空间和时间模式，空间参考和分布的信息。CSDGM是由美国联邦地理数据委员会开发制定的数字地理空间元数据标准[25]。

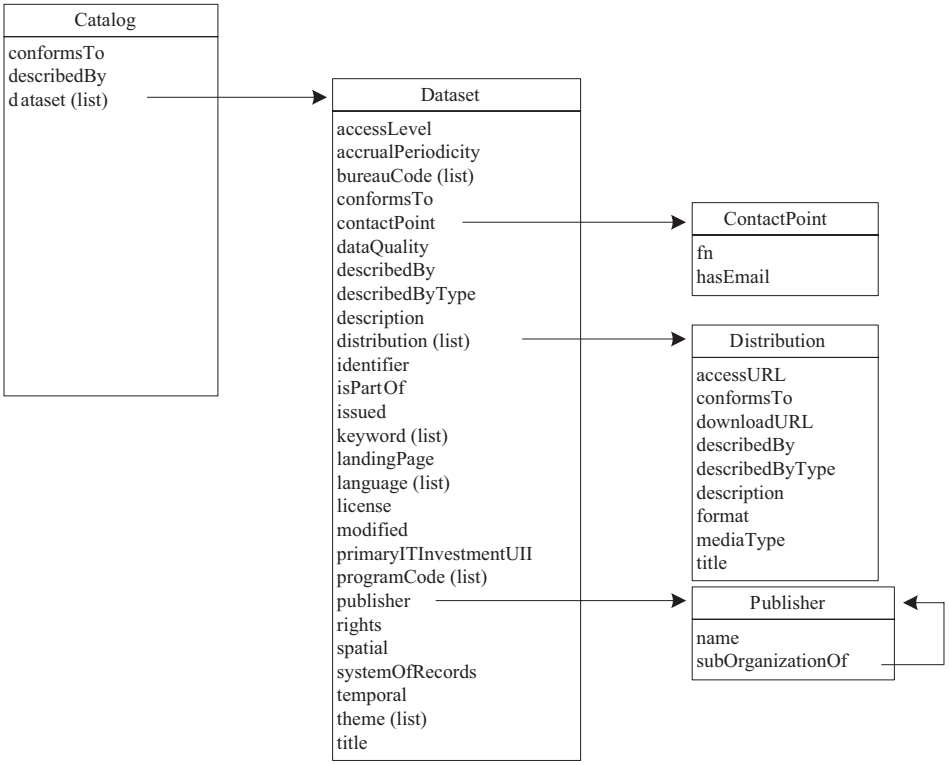

图4 POD v1.1的字段与子字段

```
"publisher": {
        "@type": "org:Organization",
        "name": "Widget Modulation Program",
        "subOrganizationOf": {
          "@type": "org:Organization",
          "name": "Office of Citizen
Services and Innovative Technologies",
          "subOrganizationOf": {
            "@type": "org:Organization",
            "name": "General Services Administration",
            "subOrganizationOf": {
              "@type": "org:Organization",
              "name": "U.S. Government"
            } } } }
```

图5 Publisher子字段实例数据的JSON片段

3.3 DCAT-AP

DCAT-AP是为提高欧盟电子政务系统语义互操作性而编写的，是以DCAT为基础，含有DCAT的所有类和属性。DCAT-AP在依据DCAT进行调整时考虑4项准则[26]。①改进发现数据集的能力。优先考虑可用于搜索和导航的方面，以及与数据集管理有关的请求，但并未考虑操作问题。②符合DCAT要求。不考虑改变DCAT模型或用其他类似元素替代DCAT元素的请求，仅添加可以从现有命名空间重用的元素，新的命名空间中不会创建新的元素。③简洁性。可以添加有证据表明是必要的元素，且元素在现有系统中通常可用。④确保应用领域中立。不考虑仅与某些特定类型数据集相关的请求。

DCAT-AP包含22个类，与POD相似，DCAT-AP也将类分为3种[27]，即强制类、推荐类和可选类。对于强制类，要求数据发送者必须提供这些类的实例信息，数据接收者必须能够处理这些类的实例信息；对于推荐类，要求数据发送者应该提供这些类的实例信息，如果信息可获得，数据发送者必须提供这些类的实例信息，数据接收者必须能够处理这些类的实例信息；对于可选类，要求数据接收者必须能够处理这些类的实例信息，数据发送者可以提供这些类的实例信息，但其没有义务提供。每个类下的属性也被分为强制属性、推荐属性和可选属性。

根据需要，DCAT-AP还复用其他词汇表。与DCAT相比，DCAT-AP增加了词汇表ADMS、OWL、Shema.mg、SPDX，具体如表2所示。

表2 DCAT-AP复用的部分词汇表

前 缀	词汇表名称	词汇表主要内容
adms	AssetDescriptionMetadataSchema	用于描述语义资产，定义高度可重用的元数据（如xml模式、通用数据模型）和参考数据（如代码列表、分类法、字典、词汇表），用于电子政务系统开发
owl	WebOntologyLanguage	旨在提供一种可用于描述网络文档和应用中所固有的那些类及其之间关系的语言
schema	Schema.org	搜索引擎（包括Bing、Google、Yahoo!、Yandex）依靠Schema.org标记改善搜索结果的显示，使用户更容易找到正确的网页
spdx	TheSoftwarePackageData Exchange	SPDX通过标准化软件供应链中许可信息的方式，帮助促进遵守免费和开源软件许可而定义的标准[28]

3.4 元数据方案的比较

POD、DCAT-AP是基于DCAT形成的，且广泛复用了其他词汇表，为数据集元数据交换提供可能，它们总体上一致，但略有区别。与DCAT相比，POD和DCAT-AP还具有3个特点。①将类和属性分为3种。POD分为必备、适用则必备和扩展（可选择），DCAT-AP分为强制、推荐和可选，二者处理本质上相同。这样做可以保证数据集元数据的简洁性和最小完整性，增加针对不同描述对象的灵活性。②添加对“关系”的描述。DCAT中缺乏对“关系”的描述。POD与DCAT-AP都增添了“关系”描述，如POD中的isPartOf字段，DCAT-AP中的dct:isVersionOf字段，通过增加“关系”属性，可以更清楚地揭示相互间联系。③添加元数据集语义关联格式的元素。如POD增添@context、@id、@type便于对数据集元数据的处理，DCAT-AP增添rdfs:Literal和rdfs:Resource。JSON-LD与RDF格式都是语义关联数据格式，可以促进数据集元数据的可发现性。

POD与DCAT-AP存在以下不同点。①POD是建立在DCAT基础上，对DCAT进行改造形成的新词汇表文件，有其自身的文本和机读格式词汇表说明。DCAT-AP是在DCAT基础上进行的扩展，将7类扩充为22类，并且包含DCAT的所有属性。DCAT-AP比DCAT更详细和复杂，描述能力也更强，更适用于多层级的数据体系。②POD添加了一些本土元素。如bureauCode和programCode都是美国独有的元素项。

4 我国政府数据开放注册管理元数据标准建设

目前，我国地方政府的政府数据开放工作已陆续展开，各地规定了相应的数据描述形式。但这些数据描述形式都不够规范，仅考虑一些基本的描述需求，且数据描述形式也不尽相同。因此，我国政府数据开放工作还缺乏统一的数据描述标准规范。为此，本研究分析相关工作中的描述需求，研究国际上通用、成熟的标准，在此基础上，提出针对我国政府数据开放和注册管理的元数据标准建设的建议。

4.1 基本结构和数据项

与国外政府数据开放情况不同，我国的政府数据开放呈现明显的多级结构。在地方政府层面进行数据开放和管理注册工作是必不可少的活动。同时，我国中央政府各机关也产生和拥有大量的政府数据，这些数据具有复杂的纵向和横向联系。因此，我国政府数据开放管理注册元数据体系应该是一个多层级的、易扩展的结构，既便于在中央政府层面进行统一管理，提高不同政府部门和机构间、政府部门与非政府机构/组织间的互操作性，又能充分考虑不同级别、不同地区的政府和非政府机构/组织在数据开放管理中的特定需求，满足未来需求的变化，实现灵活的描述体系。

为此，该规范可以在DCAT的基础上，采取类似DCAT-AP的思想和设计形式，根据描述需求对DCAT进行扩展。其中，作为强制性的核心类可包括dcat:Dataset，dcat:Catalog，foaf:Agent，dcat:Distribution，以保证数据集最基本的特征描述；数据集的主题、主题集合、开放许可文件及空间地理范围也是数据集重要的检索入口信息，因此可以作为推荐类，可以使用skos:Concept，skos:ConceptSheme，dct:LicenseDocument，dct:PeriodOfTime，dct:Location；此外，还有一些可选的数据集元数据类dct:LinguisticSystem，dact:

CatalogRecord，foaf:Document，dct:Frequency，adms:Identifier，vcard:Kind，dct:Standard，dct:RightsStatement，dct:ProvenanceStatement，dct:MediaTypeOrExtent，spdx:Checksum。这些类均来自规范的词汇表，保证了数据集跨地区、跨政府部门的可交换性。在具体应用该规范时，若增加新的类和属性，可在其基础上进行扩展。扩展的类和属性可通过复用其他词汇表或增加新的类和属性实现。

扩展时应遵循4个原则：①不能扩大类和属性的范围，以便与规范保持一致；②添加的类不能与规范中原有类相似；③添加的属性不能与规范中原有属性相似；④规范中的强制性类和属性必须保留。

上海市在采用通用元数据体系后，其元数据资源名称、摘要、数据提供方单位、数据领域、国家主体分类、部分主题分类、关键字、附件下载、首次发布日期、更新频度等数据项都通过上层通用元数据规范描述。但一些描述项，如公开属性、访问/下载次数、数据集星级评价、相关应用等在上层数据描述体系中并没有对应的类和属性，还需要通过扩展来解决。其中，公开属性可以用DCMI Metadata Terms中的dct:accessRights表示。数据集星级评价是对数据集质量的评价，POD中的dataQuality虽然是对数据集质量的评价，但主要指数据集质量是否符合数据集提供机构的信息质量标准，取值为“是”或“否”，与上海市的“数据集星级评价”含义不同，所以不能直接使用，可以定义新的数据项datasetStarRating，放在dcat:Dataset类中。“访问/下载次数”可以定义属性visit/downloadTimes，也可以定义两个属性visitTimes和downloadTimes；如果定义两个属性，visitTimes可放到dcat:Catalog类中，downloadTimes放在dcat:Dataset类中。相关应用也可以增加属性application，放在dcat:Dataset类中，由于application需要进一步对其名称、资源位置、简介等进行说明，所以可将application属性指向一个新的Application类，在Application类中再引用dct:title和dct:description等属性，或新定义其他属性。

4.2 其他词汇表的复用

为增强描述能力，同已有的描述体系保持一致，各元数据标准都尽可能复用已有的词汇表。在制定我国政府数据注册管理元数据标准时，可复用的元数据标准主要来自5个方面：①W3C推荐的元数据标准。W3C制定了很多专业化元数据标准，且已被广泛采用，上述词汇表中FOAF、SKOS等均来自于W3C。②本地相关的元数据标准。如美国在DCAT基础上加入地理信息元数据标准CSDGM。③软件工具标准。使用CKAN软件在搭建政府数据开放平台时，需要考虑CKAN的元数据元素。④元数据格式标准。美国政府数据集的注册管理元数据主要使用JSON格式，POD v1.1中的@context、@id、@type均来自JSON-LD的关键字。⑤其他元数据标准。

4.3 对已有元数据描述方案的分析

目前，我国已有多个地方政府开放数据平台，这些平台都具备各自的元数据描述方案。如北京市在对数据集描述时把资源名称、资源出版日期、资源分类、资源摘要、资源所有权单位、关键字说明、资源类型、资源记录数作为数据集元数据进行描述，而历史数据、数据下载地址、数据更新日期等也可以作为数据集目录的元数据。由于各地建立了一定的元数据集合，所以在建立统一元数据方案时应考虑相关需求。表3为北京市、上海市和武汉市数据平台的元数据描述项。

通过对表3中所列数据元数据描述项进行分析，发现主要存在以下特点。

（1）已有的地方政府开放数据描述方案仅建立简单的描述项，描述能力十分有限，只承担基本数据描述职能，甚至缺少必要描述项，如时间范围等。

（2）已有的数据描述方案基本没有规定数据取值范围。如北京市的数据描述项中“资源类型”，并没有关于“资源类型”取值的严格规定，虽然取值有“表格”，但“表格”不是一种数据格式，有CSV、XLS等不同类型格式。

（3）不同地方政府开放数据描述方案中主要的描述项基本相同，但其使用的标签并不完全一致，数据项也不尽相同。在规范的描述体系中，相同的描述项应尽量统一，而不是借助属性映射完成互操作，特殊的属性要求应在统一规范的扩展机制约束下实现。

（4）由于现有数据描述方案只考虑基本的属性描述职能，所以在设计时并没有完全参照已有规范，如DCMI。由于这些描述方案均较简单，所以在制定通用的数据描述规范时，可以方便地将已有的描述项纳入规范体系，并对现有描述数据进行转换处理。

表3 已有的地方政府开放数据元数据描述项

北京市	上海市	武汉市	DCAT相关数据项
资源名称	资源名称	资源名称	dc:title
资源摘要	摘要	数据简介	dct:description
资源所有权单位	数据提供方单位	机构名称	dct:publisher
资源分类	数据领域、国家主体分类、部分主题分类	主体分类	dcat:theme
关键字说明	关键字	关键字	dcat:keyword
数据下载地址	附件下载	资源访问	dcat:downloadURL
资源出版日期	首次发布日期	原始数据发布日期	dct:issued
数据更新日期	-	本网站更新时间	dct:modified
资源类型	-	-	dct:format
-	更新频度	-	dct:accrualPeriodicity
历史数据	公开属性	资源状态	-
资源记录数	访问/下载次数	数据条数	-
数据调用接口	数据集星级评价	机构简介	-
-	相关应用	机构地址	-

5 结语

政府数据的有效开放和充分交换需要对其进行注册管理，而元数据标准是注册管理和政府数据开放的重要基础。科学的元数据标准有利于数据集的有序组织，数据集元数据采集和交换，更有利于被数据集使用者发现和使用，提高数据集的使用效率。但是，目前我国地方政府数据开放平台的元数据标准化、规范化程度较低，数据集描述项较少，制约政府数据开放的推进和发展。本文从注册管理角度出发，研究了国外政府数据开放标准DCAT、POD和DCAT-AP基本结构和数据项，认为我国政府数据开放注册管理元数据规范应满足未来需求的变化，具有灵活的描述体系。本研究还对我国规范应包含的强制类、推荐类、可选类，以及各地方政府在具体应用该规范的扩展原则做出说明，以满足不同地方政府对数据描述的不同需求。政府数据开放注册管理元数据经过标准化、规范化建设，可进行跨国家、跨地区间的元数据交换，从而实现更大范围的政府数据共享，产生更大的经济和社会效益。

参考文献

[1] 张涵，王忠. 国外政府开放数据的比较研究 [J]. 情报杂志，2015，34（8）：142-146.

[2] 刘朝阳. 大数据定价问题分析 [J]. 图书情报知识，2016（1）：57-64.

[3] BAAS H，BROWN A. Web Services Glossary [EB/OL]. [2017-11-01]. https://www.w3.org/TR/ws-gloss/.

[4] MAALI F，ERICKSON J. Data Catalog Vocabulary（DCAT）[EB/OL]. [2018-01-16]. http://www.w3.org/TR/vocab-dcat/.

[5] W3C. DCAT Implementations [EB/OL]. [2018-01-15]. https://www.w3.org/2011/gld/wiki/DCAT_Implementations.

[6] Project Open Data Metadata Schema v1.1 [EB/OL].（2015-02-01）[2018-01-02]. https://project-open-data.cio.gov/v1.1/schema/.

[7] European Union. DCAT Application Profile for data portals in Europe Version 1.1 [EB/OL]. [2018-03-01]. https://raw.githubusercontent.com/SEMICeu/DCAT-AP/master/releases/1.1/dcat-ap_1.1.docx.

[8] European Commission. DCAT Application Profile for data portals in Europe [EB/OL]. [2018-04-11]. https://ec.europa.eu/isa2/solutions/dcat-application-profile-data-portals-europe_en.

[9] TYGEL A，AUER S，DEBATTISTA J，et al. Towards Cleaning-up Open Data Portals：A Metadata Reconciliation Approach [C] // Semantic Computing（ICSC），2016 IEEE 10th International Conference on. IEEE，2016：71-78.

[10] NEUMAIER S，UMBRICH J，POLLERES A. Automated quality assessment of metadata across open data portals [J].

Journal of Data and Information Quality (JDIQ), 2016, 8(1): 2.

[11] MARTIN S, FOULONNEAU M, TURKI S. 1-5 stars: metadata on the openness level of open data sets in Europe [C] //Research Conference on Metadata and Semantic Research. Thessaloniki, Greece: Springer, 2013: 234.

[12] BRÜMMER M, BARON C, ERMILOV I, et al. Data ID: Towards semantically rich metadata for complex datasets [M]. New York: ACM, 2014: 84.

[13] ASSAF A, TRONCY R, SENART A. HDL-Towards a Harmonized Dataset Model for Open Data Portals [EB/OL]. [2018-03-01]. https://www.researchgate.net/publication/286779372_HDL_Towards_a_Harmonized_Dataset_Model_for_Open_Data_Portals.

[14] 武琳, 黄颖茹. 开放政府数据平台元数据标准研究进展 [J]. 图书馆学研究, 2017 (6): 14.

[15] 赵蓉英, 梁志森, 段培培. 英国政府数据开放共享的元数据标准——对Data.gov.uk的调研与启示 [J]. 图书情报工作, 2016, 60 (19): 31-39.

[16] 翟军, 于梦月, 林岩. 世界主要政府开放数据元数据方案比较与启示 [J]. 图书与情报, 2017 (4): 113-121.

[17] 赵龙文, 莫荔媛, 陈明艳. 面向政府数据开放的资源描述方法 [J]. 图书情报工作, 2017, 61 (6): 115-121.

[18] Office of Management And Budget. max. gov Home.page [EB/OL]. [2017-12-23]. https://max.gov/maxportal/home.action.

[19] 黄如花, 林焱. 国外开放政府数据描述规范的调查与分析 [J]. 图书情报工作, 2017, 61 (20): 37.

[20] GSA. MAX. gov Shared Services [EB/OL]. [2018-01-23]. https://apps.gov/products/max-shared/.

[21] CKAN. Metadata [EB/OL]. [2018-02-06]. https://ckan.org/portfolio/metadata/.

[22] Project Open Data. Metadata Resources for Schema v1.1 [EB/OL]. [2018-01-01]. https://project-open-data.cio.gov/v1.1/metadata-resources/#field-mappings.

[23] Dan Brickley. About Schema.org [EB/OL]. [2018-02-04]. http://schema.org/docs/about.html.

[24] International Organization for Standardization. ISO 19115: 2003 [EB/OL]. [2018-02-04]. https://www.iso.org/standard/26020.html.

[25] Geospatial Metadata [EB/OL]. [2018-02-05]. https://www.fgdc.gov/metadata.

[26] Joinup. DCAT application profile for data portals in Europe [EB/OL]. [2018-02-05]. https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe/releases.

[27] Joinup. DCAT DCAT-AP v1.1. [EB/OL]. [2018-02-05]. https://joinup.ec.europa.eu/release/dcat-ap-v11.

[28] SPDX. About [EB/OL]. [2018-02-03]. https://spdx.org/about. Towards_a_Harmonized_Dataset_Model_for_Open_Data_Portals.

作者简介

赵盼云，女，1992年生，硕士研究生，研究方向：政府数据、关联数据、知识图谱，E-mail：flyzpy@qq.com。
耿骞，男，1965年生，博士，教授，研究方向：信息检索、网络信息管理、管理信息系统，E-mail：gengqian@bnu.edu.cn。
柏林，女，1992年生，硕士研究生，研究方向：信息分析，E-mail：bailin0802@bnu.edu.cn。

Research on Open Government Data Registration Management Metadata

ZHAO PanYun GENG Qian BAI Lin
(School of Government, Beijing Normal University, Beijing 100875, China)

Abstract: In order to solve the problem of unclear definition of open datasets, untimely datasets updating and lack of effective data exchange mechanism between government departments, government departments and social organizations in the opening of government data. This paper studies the open registration management of government data, the related elements and registration process in registration management, and on this basis, discussed the metadata standards closely related to data registration. Through the analysis of the description requirements of the datasets in the registration management process, and combined with the analysis of the foreign government data open standards DCAT, POD and DCAT-AP, the paper proposes the construction of the metadata standard for the open registration management of government data in China.

Keywords: Open Government Data; Registration Management; Metadata Standard

（收稿日期：2018-04-19）