关联数据质量词表及其应用研究*

贾君枝
（中国人民大学信息资源管理学院，北京 100872）

摘要：关联数据集的重用与消费是关联数据应用的重要部分，如何获得数据质量信息并进行有效的评估成为关键问题。本文旨在分析不同参与主体可能产生的数据质量信息，引入数据质量词表，对其特征进行分析，并讨论RDF序列化，以客观地记载这些质量信息，最终形成关于数据质量事实链，实现数据的追踪与利用。

关键词：数据质量；词表；关联数据
中图分类号：G254
DOI：10.3772/j.issn.1673-2286.2018.12.002

语义网技术的发展推动关联数据集不断出现，关联开放数据云（linked open data，LOD）项目2018年6月的数据集已达1 224个，链接数16 113个[1]。随着数据集的增长，数据的重用、消费不断发生，而数据质量在一定程度上会影响用户的使用效率，判断数据质量成为数据消费之前的一个重要决策，数据质量评估应运而生。由于不同的评估主体参与、采用的评估指标体系差异，单个机构的评估结果并不完全可信，而且这些评估结果并没有伴随数据集而存在，导致用户获得质量评估数据的难度增加。因此，有效地记录数据质量的不同维度信息，调动多个机构参与到数据质量评估建设中，对于数据消费者而言将可能获得各个层面的数据质量信息，为其数据集的选择判断以及对于数据发布者及开发者的信任度提供充分的数据支持。因此有效地描述数据集在不同阶段产生的各种类型的数据质量信息，对于数据生产者、开发者、消费者各个主体而言，都具有重要价值。

基于此，2004年2月W3C专利政策运营小组编制数据质量词表，并由Web最佳实践工作组发布数据。该词汇表旨在使发布、交换、消费高质量元数据变得更为容易，能够记录数据各个生命周期阶段关于数据质量的元数据信息，以帮助用户进行有效的选择与判断。本文旨在研究不同参与主体可能产生的数据质量信息，通过数据质量词表来客观地记载这些信息，以形成关于数据质量事实链，实现数据的追踪与利用。

2009年，Berners-Lee[2]提出关联数据的“五星标准”，旨在提高关联数据质量。随后Hoxha等[3]提出“绿色关联数据”的原则，魏来等[4]基于“绿色关联数据”总结出包括内容、表述、系统与应用的关联数据质量标准总框架。德国莱比锡大学Zaveri等[5]提出针对关联开放数据云图的数据质量评估框架。王振蒙等[6]利用RDF词汇、URI有效性等指标对5家国家图书馆发布的图书关联数据集进行分析和评估。Wei等[7]认为目前缺乏基于数据质量词表的数据集质量标注工具的系统研究，提出可视化用户接口以实现对数据集的质量标注。可以看出，关联数据质量评价日益得到重视，但如何有效地描述数据质量评价结果并未在相关研究中提及。

1 不同参与主体所涉及的数据质量分析

数据集的数据质量形成源于不同主体的共同参与，这些参与主体有数据提供者、数据质量评价者、数据消费者。除这些主体外，数据质量形成还依赖于一定的数据质量政策与法律。

*本研究得到国家社会科学基金重点项目“基于关联数据的中文名称规范档语义描述及数据聚合研究”（编号：15ATQ004）资助。

1.1 数据提供者

数据提供者指收集、发布数据的机构，负责数据的更新与维护管理，旨在提高其声誉及社会参与度。数据提供者作为数据来源机构，由于各机构所提供的数据量较大且结构各异，对数据本身的理解、描述存在偏差，所发布的数据源可能存在不一致、相互矛盾及冲突。显而易见，关联数据同一般数据相比，具有自身的特点，其质量主要取决于数据提供者，质量层面包含的主要内容有数据集元数据信息（数据集的大小、发布机构、主题等）、数据模型（类、属性定义）、数据格式、数据发布状况（如发布时间、更新频率、是否接受用户修改等）、数据获取方式等。

1.2 数据质量评价者

数据质量评价者对所发布的数据集进行评价，获得的评价结果可以为数据消费者提供选择参考。通常数据质量评价者选择要评价的数据集须依据一定的数据质量评估体系，对各指标进行量化计算以获得可信度高的结论。数据质量评价者包括第三方评价机构、个人，通过对各类数据集的评价打分、排名，旨在获得高质量的数据集供用户选择使用。评价过程中，选用不同的数据质量评估体系决定评估的结果。ISO/IEC 25012将数据质量分为内在质量和系统依赖质量两大维度[8]，共包括15个指标。内在质量有准确性、完整性、一致性、可信度、现时性，系统依赖质量有可检索性、准确率、保密性、效率、遵从性、可用性、可理解、可追踪、可携带、可恢复性。针对关联数据的数据质量评估体系，目前较权威的是Zaveri等[5]提出的指标体系，且将其分为存取性、内在性、上下文、表示四大维度。存取性包括可用性、授权、链接、安全性、性能，内在性包括语法验证、语义准确、一致性、简洁性、完整性，上下文包括相关性、可信度、可理解性、及时性，表示包括简洁性、互操作性、可解释性、可视化。

1.3 数据消费者

数据消费者不仅浏览数据，而且贡献、提供质量反馈。他们可以编辑数据，实现数据的纠错及其更新，有助于数据的维护；此外，他们也可对数据提供评论、标注。通过有效的反馈环节使数据质量趋于完善。数据消费者是关联数据集的最佳实践者，作为数据集的使用对象，他们有权且最有资格对数据质量进行评价监督，通过对数据质量反馈信息以保证数据质量处于不断上升状态。有效地设置由数据消费者参与的关联数据质量反馈环节，采用多手段积极倡导消费者参与，及时地搜集用户反馈信息，将有助于关联数据的质量提升。

2 数据质量词表的框架结构

关联数据集自身并不包括对其数据质量的描述，有效地记录不同主体所参与的数据质量活动，将成为获得数据质量、追踪数据质量的重要依据。为保证描述记录的可理解性、可操作性，需要制定专门型词表对其表示。因此，W3C的Web数据最佳实践工作组于2016年正式发布数据质量词表（data quality vocabulary）[9]，其有机地记录数据的评估过程及其结果，反映了用户反馈。

2.1 数据质量词表的特征

2.1.1 数据目录词表的扩展

数据目录词表（data catalog vocabulary，DCAT）作为网络数据目录的互操作词表，旨在实现不同格式的数据共享与交换[10]。DCAT主要用于表示政府数据目录，定义了3个基本类，即目录、数据集、发布方式。目录定义了目录名、发布者、时间、地点、语种、所包含的数据集，数据集定义了数据集名称、发布者、关键词、描述、时间、地点、语种，发布方式定义了数据集的授权、存取URL、类型、格式、大小。DCAT词表对数据集的基本发布状况进行准确且较全面的描述，为数据质量词表的制定奠定了基础。但数据质量词表只侧重于对数据质量进行描述，关于描述对象数据集本身则应用DCAT词表，其所定义的数据质量类与DCAT词表中的目录、数据集、发布方式息息相关，实际是对数据集的质量元数据进行描述。

2.1.2 数据质量描述框架的确立

数据质量描述词表提供了数据质量描述框架，定义了数据质量描述中所涉及的类、属性、实例，构建了不同数据集质量描述的概念模型，为基于数据质量的各种应用提供可能。从其描述内容看，定义了数据质量

评价对象、评价所采用的指标体系、评价结果值、评价政策及遵循的标准、用户反馈等，对数据质量评价过程及评估方法进行准确记载，并明确各个实体类间关系，以鼓励不同人员参与数据质量评价，全面地反映评价者的观点、评注及其相关证据。这些有助于帮助数据消费者进行选择判断。但是，其并不关注数据本身的质量问题，不对数据质量进行评价；其旨在实现用户及机器对这些质量数据的解读，有助于用户对数据集进行标注、评价、比较、选择，追踪数据质量的动态变化状况，为后期数据集成应用提供参考。

2.1.3 重用其他词表

数据质量词表构建并不是从零开始，而是在充分吸收现有多个词表的基础上发展而来，以实现最小成本构建。各个词表共同表述数据集质量信息，相互补充构成对数据集质量活动的完整描述。因此，除了定义自身特定的类及属性（命名空间定义为dqv），其重用了其他本体的类及属性作为描述构成。重用的本体有数据目录词表（DCAT）、都柏林核心元素集（DCMI）[11]、数据集使用词表（DUV）[12]、简单知识组织系统（SKOS）[13]、数据起源（PROV）[14]、Web注释词表（OA）[15]、ODRL词表[16]、数据立方体词表（QB）[17]等。数据目录词表用于定义数据集的特征信息，明确数据集对象。都柏林核心元素集用于描述通用类型的数据，如数据集的标题、数据标准。数据集使用词表描述了消费者关于数据集的使用经验、引用及其反馈信息，定义了评价反馈、使用、使用反馈、使用工具等基本类。数据起源描述了数据集产生、修改、拥有及其他影响的元数据，定义了实体、活动、代理3个基本类，用以追踪对数据集所产生影响的人员、活动及变化，如数据质量标准与评估体系之间的使用及生成关系采用此定义。简单知识组织系统定义了共享与链接知识组织系统的模型，提供了知识组织系统中概念及概念之间关系、不同词表映射的表示词汇。ODRL词表旨在发展促进开放式国际政策语言表述，支持发布、分配、消费内容、应用及服务中数字资产的透明且创新式使用；涉及政策类型，允许、禁止的职责行为，所扮演的功能角色、数字资产关系。Web注释词表定义了有效表达标注行为的互操作框架，用来描述关联数据环境下用户对网络数据的评注行为，客观记录评注人对评注对象所实施的评论、选择等活动。数据立方体词表用于交换及共享统计数据及元数据。

数据质量词表只定义了自身的核心类（如质量评估、质量标注、用户质量反馈、数据质量元数据），其他类都来源于其他词表。同时其将核心类通过子类、子属性关系与其他词表建立联系，如数据质量标注类放于OA词表的标注类下，评估的结果放于QB的数据集类下，这些为实现多个词表的互操作提供了可能，旨在充分发挥数据网络的价值。

2.2 数据质量词表的结构

数据质量词表实施的评估对象是数据集，主要记录对数据集所开展的质量评估、标注、元数据等一系列质量管理活动。

2.2.1 数据质量评估

数据质量评估需要明确所制定的数据质量政策、采纳的数据质量标准及其所应用的数据质量评价指标体系。数据质量政策指导数据质量活动，为其提供行动准则，通常包含目标、背景、范围、角色及职责、政策声明及定义。数据质量标准是保证数据质量管理活动具有可控性的重要手段，旨在形成跨国家、组织的统一性数据质量管理方法，以实现数据存储、传递和共享，促使各评估机构遵循统一的数据质量评估标准，在一定程度上降低数据质量评估成本。数据质量标准通常定义满足数据质量需求的一系列特征，对其进行解释说明并分层展示，实际上为数据质量评价提供指标体系。数据质量评估是依据数据质量政策及标准而实施的评估过程，以明确获得评价结果，评估过程涉及评估对象、评估指标（定义数据结构）及结果值。数据质量词表定义了3个基本大类，即质量政策（dqv:QualityPolicy）、标准（dcterms:Standard）、评估（dqv:QualityMeasurement）。评估指标体系又细分为3个子类：类（Category）、维度（Dimension）、指标（Metric），类划分为若干维度，维度下细分为若干指标。

2.2.2 数据质量标注活动

标注是创建不同资源之间的关联行为，数据质量标注旨在表达数据资源与资源的关系信息，一个完整的标注情境包括标注者、标注对象、标注行为、标注

内容、时间。标注对象实际为标注目标，标注内容称为标注主体，表达对目标的标注内容。标注行为包括评价（评语、评级）、标签、收藏、描述、提问、回答、识别、分类、描述、编辑、联接、加亮等活动。数据质量标注描述了数据质量认证及反馈信息，包括数据质量认证、用户质量反馈。ISO在ISO/IEC指南2中将质量认证定义为第三方依据程序对产品、过程或服务符合规定的要求给出书面保证（合格证书）[18]。数据质量认证是第三方对数据质量符合质量标准给出书面保证，包含数据质量标准体系和数据集质量认证。数据质量认证将数据集与证书之间建立关联，数据集称为对象，证书称为主体，通过评估过程建立两者联系。用户反馈是识别用户需求、评估用户满意度、发现质量问题的重要方法[19]。用户质量反馈从用户角度反映对数据集的满意程度，通过用户参与来提供数据质量，包含用户、反馈方式、反馈内容。反馈方式来自标注行为类型，有评级、评语、提问、分类、描述、编辑等；反馈内容涉及内容主体及所提及对象。数据质量词表定义了数据质量标注类（dqv:QualityAnnotation），其划分为两个子类，即质量认证类（dqv:QualityCertificate）、用户质量反馈类（dqv:UserQualityFeedback）。

2.2.3 数据质量元数据

数据质量元数据描述了数据质量的基本信息，有助于用户迅速获取数据集的基本质量信息，指导用户进行查询及使用。数据质量元数据已成为数据质量的基本构成，包含数据源本身、数据质量认证、政策、数据集的评估活动、标注的记录。数据质量词表定义了质量元数据类（dqv:QualityMetadata）。

如图1所示，描述完成这3个基本活动后，则构建类之间的关系，形成数据模型[9]。

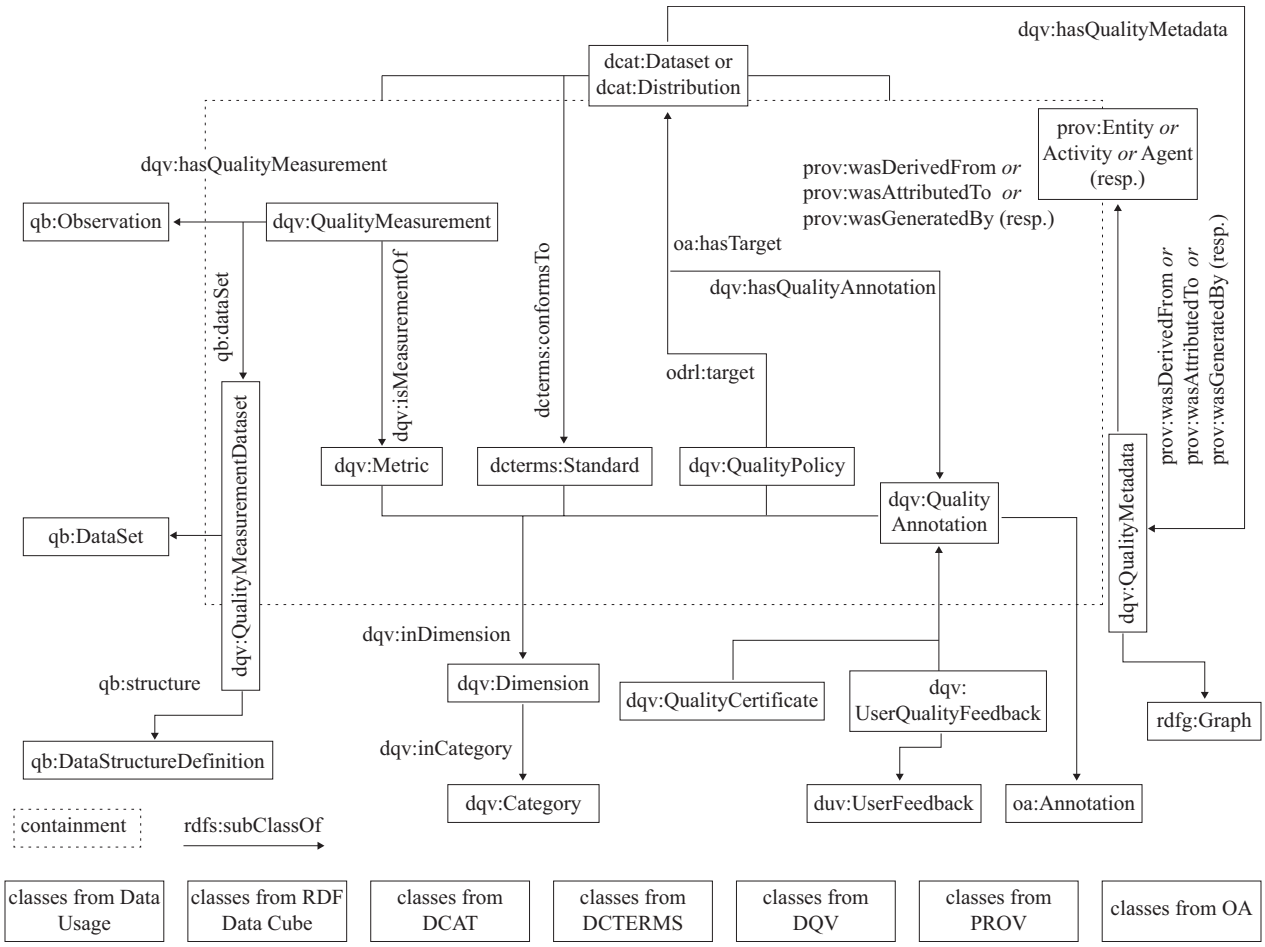

图1 数据质量词表的数据模型

3 数据质量词表的RDF描述应用

应用数据质量词表可以准确地实现对数据质量评估、标注及元数据信息进行描述，据此用户或机器可以及时获取数据质量信息，为数据的消费及再利用提供依据。BNB是大英图书馆发布的RDF/XML格式的关联书目数据集，其包含图书、期刊、报纸等图书馆收藏的资源。大英图书馆的BNB数据集作为较早发布关联书目集的国家机构，成为许多机构所选用的数据集评价对象。现选用其图书子集进行RDF描述，利用一定的评估指标及其用户标注行为对其进行综合评价，以展示该数据集部分质量情况。

3.1 数据质量评估的RDF

当前选用Zaveri等[5]提出的指标体系（https://www.w3.org/2016/05/ldqd，命名空间为ldqd）对BNB数据集（http://bnb.data.bl.uk）的图书子集进行评估，对可用性指标进行评估，结果表明该数据集URL可以被访问。

```
Dataset1a dcat:Dataset ; # 数据集对象
        dcterms:title"BNB LOD Books";
        dct:issued"2018-10-01"^^xsd:date;
        dct:publisher:The British Library;
        dct:language<http://id.loc.gov/vocabulary/iso639-1/en>;
        dcat:byteSize"1,354,076"^^xsd:decimal;
Dataset1 dqv:hasQualityMeasurement:measurement1;# 数据集评估结果
ldqd a dcterms:standard;
measurement1 a dqv:QualityMeasurement;
        dqv:computedOn Dataset1;
        Dcterms:conFormsTo ldqd;
        dqv:isMeasurementOf :downloadURLAvailabilityMetric;
        dqv:value"true"^^xsd:boolean;
downloadURLAvailabilityMetric    # 数据集采用的评估指标
        a dqv:Metric;
        skos:definition"It checks if dcat:downloadURL is available and if its value is
        dereferenceable."@en;
        dqv:expectedDataType xsd:boolean;
        dqv:inDimension:availability;
availability
        a dqv:Dimension;
        skos:prefLabel"Availability"@en;
        skos:definition"Availability of a dataset is the extent to which data (or some portion of it) is present,obtainable and ready for use."@en;
        dqv:inCategory ldqd:accessibility Dimensions;
ldqd:accessibilityDimensions a dqv:Category;
        skos:prefLabel"Accessibility"@en;
```

3.2 数据标注的RDF

用户对BNB的图书子集的可用性进行评级，给予四星级分值。

```
Dateset1Classfication a oa:Annotation;  # 用户给定四星级评分
        a dqv:UserQualityFeedback;
        oa:hasTarget Dataset1;
        oa:hasBody four_stars;
        oa:motivatedBy dqv:qualityAssessment, oa:classifying;
        dqv:inDimension :availability.
four_stars                  # 四星级的定义
        a skos:Concept;
        skos:inScheme:OpenData5Star;
        skos:prefLabel"Four stars"@en;
        skos:definition"Dataset available on the Web with structured machine-readable non
        proprietary format. It uses URIs to denote things."@en.
```

3.3 数据质量元数据的RDF

对大英图书馆的图书数据子集进行的评估及其标注活动的元数据信息进行描述。

```
Dateset1Metadata a dqv:QualityMetadata;  # 质量元数据
        prov:generatedAtTime"2018-11-11"^^xsd:date;
        prov:wasGeneratedBy:qualityEvaluation.
        prov:generatedAtTime"2018-11-12"^^xsd:date;
        prov:wasGeneratedBy:qualityAnnotation.
qualityEvaluation           # 评估元数据
        a prov:Activity;
        rdfs:label"The evaluation of Dataset1's quality"^^xsd:string;
        prov:used           Dataset1;
        prov:generated      Dateset1Metadata;
        prov:endedAtTime    "2018-11-11"^^xsd:date;
```

```
        prov:startedAtTime      "2018-11-11"^^xsd:date;
qualityAnnotation               # 标注元数据
        a prov:Activity;
        rdfs:label"The evaluation of Dataset1's quality"^^xsd:
        string;
        prov:used               Dataset1;
        prov:generated          Dateset1Metadata;
        prov:endedAtTime        "2018-11-12"^^xsd:date;
        prov:startedAtTime      "2018-11-12"^^xsd:date.
```

通过对该数据集的质量评估过程的描述，可以清晰地展示其评估中所采用的指标体系及其评估结果，并充分地表示了用户所参与的评估活动类型及其标注内容，这些有助于数据消费者在后期选择使用该数据集时，形成基于数据质量评估活动的一系列准确决策。

4 结语

随着开放关联数据集的增长，数据质量成为消费者关心的重要问题。本研究对影响数据质量的各种活动进行阐述，深入地对W3C发布的数据质量词表的特征进行细致深入的分析，并对其RDF应用场景进行说明。随着该词表的不断普及应用，越来越多的机构及用户参与到数据质量的相关描述中，在未来将极大地推动数据质量的提升，真正发挥数据价值作用。

参考文献

[1] Insight Centre for Data Analytics. The Linked Open Data Cloud [EB/OL]. [2018-10-30]. https://lod-cloud.net/.

[2] BERNERS-LEE T. Linked Data [EB/OL]. (2009-06-18) [2018-10-30]. https://www.w3.org/DesignIssues/LinkedData.html.

[3] HOXHA J, RULA A, EII B. Towards green linked data [EB/OL]. [2018-10-30]. https://www.researchgate.net/publication/228430782_Towards_Green_Linked_Data.

[4] 魏来，付瑶. 基于greenlinkeddata的关联数据质量标准 [J]. 情报资料工作，2013，34 (3)：70-73.

[5] ZAVERI A, RULA A, MAURINO A, et al. Quality assessment for linked data: A survey [J]. Semantic Web, 2016, 7 (1): 63-93.

[6] 王振蒙，姜恩波. 关联书目数据质量评估框架构建与实证评估 [J]. 图书情报工作，2016 (15)：108-115.

[7] WEI J, XU Z M, XIA W Z. DQAF: Towards DQV-Based Dataset Quality Annotation Using the Web Annotation Data Model [C]//Web Information Systems & Applications Conference. IEEE, 2017.

[8] ISO/IEC 25012 [EB/OL]. [2018-09-13]. http://iso25000.com/index.php/en/iso-25000-standards/iso-25012.

[9] W3C Working Group. Data on the Web Best Practices: Data Quality Vocabulary [EB/OL]. (2016-12-15) [2018-09-13]. https://www.w3.org/TR/vocab-dqv/#DimensionsofZaveri.

[10] Data Catalog Vocabulary (DCAT) [EB/OL]. (2014-01-16) [2018-09-13]. https://www.w3.org/TR/vocab-dcat.

[11] Dublin Core Metadata Initiative [EB/OL]. [2018-09-13]. http://dublincore.org/documents/dcmi-terms/.

[12] World Wide Web Consortium. Dataset Usage Vocabulary [EB/OL]. [2018-09-13]. https://www.w3.org/ns/duv#.

[13] MILES A, BECHHOFER S. SKOS Simple Knowledge Organization System Reference [EB/OL]. (2009-08-18) [2018-09-13]. https://www.w3.org/TR/skos-reference.

[14] LEBO T, SAHOO S, MCGUINNESS D. PROV-O: The PROV Ontology [EB/OL]. (2013-04-30) [2018-09-13]. https://www.w3.org/TR/prov-o.

[15] SANDERSON R, CICCARESE P, YOUNG B. Web Annotation Vocabulary [EB/OL]. (2016-06-05) [2018-09-13]. https://www.w3.org/TR/annotation-vocab/.

[16] IANNELLA R, GUTH S. ODRL Version 2.1 Common Vocabulary [EB/OL]. (2015-03-05) [2018-09-13]. https://www.w3.org/community/odrl/model/2.1/.

[17] CYGANIAK R, REYNOLDS D. The RDF Data Cube Vocabulary [EB/OL]. (2014-01-16) [2018-09-13]. https://www.w3.org/TR/vocab-data-cube/.

[18] ISO/IEC Guide 2: 2004. Standardization and related activities—General vocabulary [EB/OL]. [2018-09-13]. https://www.iso.org/iso/catalogue_detail?csnumber=39976.

[19] FOTROUSI F, FRICKER S, FIEDLER M. Quality Requirements Elicitation Based on Inquiry of Quality-Impact Relationships [C]//IEEE 22nd International Requirements Engineering. Karlskrona, 2014.

作者简介

贾君枝，女，1972年生，博士，教授，博士生导师，研究方向：信息组织，E-mail：junzhij@163.com。

Construction and Application of Quality of Linked Data Vocabulary

JIA JunZhi
(School of Information Resource Management, Renmin University of China, Beijing 100872, China)

Abstract: Reuse and consumption of linked data sets is an important part of the application of linked data. How to obtain data quality information and make effective evaluation becomes a key problem. This paper aims to analyze the data quality information that may be generated by different participants, introduces data quality vocabulary, analyzes its characteristics, and discusses RDF serialization to objectively record these quality information, so as to form a chain of facts about data quality and realize data tracking and utilization.

Keywords: Data Quality; Vocabulary; Linked Data

（收稿日期：2018-11-15）

书 讯

《中国高被引分析报告2017》

《中国高被引分析报告2017》按理、工、农、医、人文、社科等领域划分为50个学科，综合分析各个学科的高影响力论文、研究热点与前沿、高影响力期刊、高影响力作者和高影响力科研机构，并以关联图谱的方式展现了多种学术关系，有助于科研人员及时发现并跟踪研究热点，有利于期刊编辑部监测本刊学术影响力，有利于科研管理机构评估科研能力，是高等院校、科研院所及期刊编辑部等相关单位和人员的参考工具书。

该书以“中国知识链接数据库”为依托，数据覆盖我国6 000余种期刊的论文及引文。书中分学科揭示了高影响力的学者、研究机构（大学、研究所、医院等）、地区（省/自治区/直辖市）、学术期刊、图书、外文期刊和会议录，并采用共词分析、共被引分析和合著分析等方法绘制出各学科的前沿主题分布以及作者、机构和期刊间关联的知识图谱。

2014—2017年的《中国高被引分析报告》均由中国科学技术信息研究所编制，曾建勋主编，科学技术文献出版社出版。欢迎业界同仁鉴阅订购。