面向互操作性的捷克开放数据质量评估

Dasa Kusnirakova[1], Mouzhi Ge[2],Leonard Walletzky[1],和 Barbora Buhnova[1] 信息学院，马萨里克大学，布尔诺，捷克共和国[2]代根多夫应用技术大学，代根多夫，德国{*kusnirakova, wallet, buhnova*}*@mail.muni.cz, mouzhi.ge@th-deg.de*

关键词: Open 数据、数据质量、数据互操作性、评估框架 架

摘要: 随着已发布开放数据集数量的快速增长，在考虑开放数据质量的同时，支持智慧城市中的开放数据发展至关重要。在捷克共和国及其国家开放数据目录（NODC）中，开放数据集通常仅根据其元数据进行评估，而将内容以及对推荐数据结构的遵循情况完全交由数据提供者负责。开放数据集的互操作性仍然未知。因此，本文旨在提出一种新颖的内容感知质量评估框架，根据五个数据质量维度评估开放数据集的质量。借助所提出的框架，我们对发布于 NODC 的捷克开放数据集的面向互操作性的数据质量进行了基础性分析。评估结果表明，特定领域开放数据质量评估能够检测出超越传统启发式方法所发现的捷克开放数据质量问题，提高其互操作性，从而增强其为社会创造价值的潜力。这项研究的发现不仅适用于捷克共和国，也可应用于其他希望改进开放数据质量评估流程的国家。

arXiv:2303.01950v1 [cs.DB] 3 Mar 2023

1 引言

随着开放数据范式在全球范围内得到更广泛的采用，已发布数据集的数量不断增加，处理数据的重点也随之转移，数据质量已成为组织关注的主要问题（Sadiq 和 Indulska，2017）。数据质量的不确定性正变得愈发关键，因为低质量数据集创造附加价值的能力非常有限，有时甚至会损害应用程序和服务（Ge 和 Lewoniewski，2020）。

关键问题之一是开放数据集的互操作性，因为开放数据的概念基于可重用性以及与其他数据互联的理念（Bizer 等，2011）。由于数据来自不同的发布者，其结构或单个值往往彼此不兼容（Ge 等，2019）。数据集的这种多样性（例如所用数据格式或词汇表各异）会显著增加后续数据使用所需的处理工作量（Thereaux，2020），甚至可能使数据合并完全无法实现。

因此，本文旨在提出一种新颖的质量评估模型，从五个主要数据质量特征来衡量开放数据集的质量。该模型将数据集的互操作性作为优先考量，评估框架着重关注数据集的互操作性，同时兼顾内容和数据结构。此外，我们展示了对旅游领域捷克数据集进行面向互操作性的数据质量评估所得的见解。这些数据集发布于捷克国家开放数据目录(NODC)。最后，我们认为，面向特定领域且面向互操作性的开放数据质量评估，除了常用于评估捷克开放数据质量的技术之外，还能够识别出多个严重的数据质量问题。

2 相关工作

近年来，针对开放数据质量问题的研究取得了进展。例如，在（Berners-Lee，2012）中，作者提出了五星开放数据评级系统。经过质量分析的数据会根据既定的质量要求获得相应的星级。尽管这一工具在欧洲得到广泛使用，并由欧洲数据门户推广（Carrara 等，2018），但其结果可能会产生误导。该评估仅关注数据质量的部分维度（主要是法律和技术方面），

因此可能无法反映用户对数据质量的所有需求。

大多数已发表的论文都研究开放数据门户。在（Ubaldi，2013）和（Viscusi 等，2014）中，作者对开放政府数据的质量评估进行了深入研究。他们创建了一系列指标，根据数据的可用性、需求、再利用、格式或时效性来判定数据质量。然而，与此前的研究类似，他们提出的指标都未考虑数据集的内容；所有指标仅在数据集或门户层面发挥作用。

评估其中 数据文件中的开放数据质量并非全新想法。早期基于互操作性（例如时效性或完整性）的评估指标之一由（Vetr`o 等，2016）提出。尽管该论文提出了测量粒度最细的质量维度，但其框架缺乏对所考察数据的语法和语义评估。（Nikiforova，2020）则提出了一种更通用的方法，即通过可执行评估模型支持自定义数据质量规范。 然而，该方法依赖于对数据质量要求进行精确、正确的规范定义，并需要大量人工干预。

3 开放式正式标准

捷克共和国引入了一种旨在确保开放数据集互操作性的标准化技术，称为开放式正式标准（ OFS）。开放格式标准是针对特定领域制定的技术指南，由捷克内政部协调，通过协作式决策流程开发。其目标是简化数据使用，并确保数据互操作性，即使不同的数据提供者提供相同类型的数据也是如此。在技术、语法和语义维度上，均可确保完全互操作性（OHA，2021）。

确保互操作性是发布者在发布与开放式正式标准所建模内容相关的数据时应遵循这些标准的主要原因。此外，根据《第106/1999号信息自由获取法》，开放式正式标准对开放数据发布者具有约束力。

每个 OFS 都从对特定数据集基本术语的描述开始，这些术语统一了数据的语义，也就是数据的理解方式。这些术语以概念方案的形式呈现，将术语建模为类、类的属性以及它们之间的关系（Kl´ ımek，2020a）。

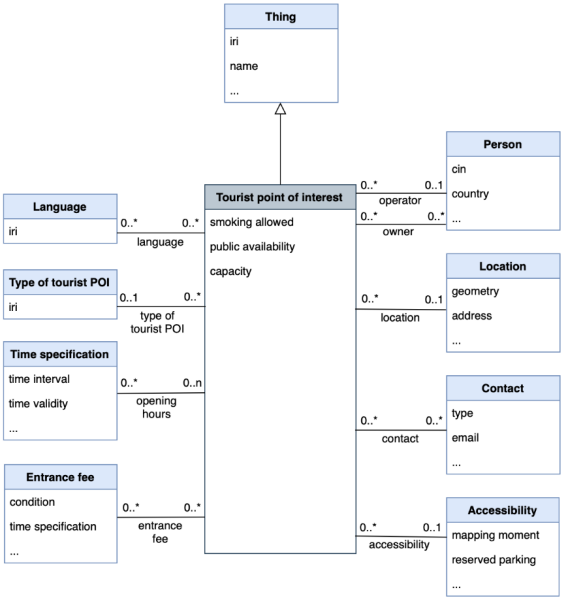


图1：旅游兴趣点对象的开放式正式标准方案示例（ MV ˇCR，2020）。译为英语。

方案也以图形方式呈现，便于理解，如图 1 所示。所有已发布的 OFS 均采用统一的格式。

OFS 的一个关键特点是，出现在多个开放格式标准中的类（例如 人物、联系人 或位置）都在同一处进行定义，即所谓的共享规范。共享规范确保不同数据集中相同实体之间的兼容性，从而便于按照各种开放格式标准处理数据。例如，遵循事件 OFS 的数据集中所提供的音乐会入场信息，与旅游兴趣点数据集中发布的城堡入场信息采用相同的表示方式。

3.1 开放格式标准采用的现状

捷克国家开放数据目录不检查各个数据集的质量；NODC 仅处理相应数据文件的元数据（Kl´ı mek，2019）。尽管内政部通过各种工具支持数据提供者提升数据质量，例如定期评估数据集的元数据，或提供常见不良实践示例（Kl´ımek，2022），但各个数据集面向互操作性的数据质量完全由数据提供者负责（V´ıta，2021）。

尽管 OFS 展现出巨大潜力，且对数据发布者具有法律约束力，但其在实践中的实际应用并不理想。我们检查了

目前对旅游兴趣点 OFSs 的遵循情况，方法是使用开放数据国家目录 API 手动检索：我们搜索了在数据集元数据的任意属性值中包含旅游兴趣点标准标识符的数据集，这一要求由 OFS 规定。结果令人意外。迄今为止，尚无任何关于这一主题的已发布数据集遵循该标准。这些发现促使我们进一步研究这一问题。

4 捷克开放数据质量评估框架

为设计评估框架，我们采用了 Wang 和 Strong（1996）提出的方法。所提出的数据质量框架旨在理解哪些特征与捷克开放数据相关，同时确保即使是普通公众也能轻松理解评估结果和指标。评估框架的总体特征如下所述。

4.1 分数计算

所提出框架的每个维度都会根据数据集在所考察方面的质量，为其赋予一定数量的点数。得分范围为 0 至 100 分，得分越高，数据集的质量越高。各个维度计算出的得分彼此独立，不会合并为一个最终得分。其目的是突出数据集在互操作性以及遵循特定 OFS 方面的优势和缺陷。

4.2 特征

每个 OFS 通常都包含一个最小示例，其中规定了对建模实体的最低要求。如果数据集包含的信息少于最小示例中提供的信息，那么这些数据很可能毫无意义，也不会有人能够使用它（Klˊımek，2020b）。因此，尽管根据官方文档（Dvoˇrˊak 等，2020），OFS 模式中规定的所有实体都是可选的，但所提出的评估框架仍将这些最低要求（称为特征）视为强制性 要求。

除了最小示例之外，开放格式标准还提供了一个更复杂的示例。这种复杂模型描绘了应当填写的元素，以便向用户提供有关该模-型域最准确的图景。所提出的框架将附加特征视为可选 内容。

4.3 特征权重

由于每个特征的重要性各不相同，因此它们对所考察维度最终得分的影响也应有所不同。缺少必选特征（例如旅游目的地的位置）会在数据处理阶段造成远比缺少可选特征（例如目的地使用的语言）严重的问题。单个特征的这种重要性称为特征权重。

特征权重由特征类型决定。必选特征的权重显著更高，因为它们承载着最关键的信息；缺少此类特征会大幅降低数据集的质量。具体而言，所提出的框架规定，一个必选特征的权重等于所有可选特征权重之和。由于模型采用 0 到 100 的量表，因此权重计算如下：

$$\Sigma\, w_{mf} + \Sigma\, w_{of} = 100 \tag{1}$$

$$w_{mf} = \frac{100}{N_{mf} + 1} \tag{2}$$

$$w_{of} = \frac{w_{mf}}{N_{of}} \tag{3}$$

其中，w表示特征权重，N 表示特征数量，mf 表示必选特征，of 表示可选特征。

5 数据质量维度

在整个开发过程中，我们识别出十个质量维度，其中选出的这五个似乎是评估数据集以下方面最明确的候选项：a）互操作性，b）对开放格式标准所定义规则的遵循情况，以及 c）基本质量方面；而这些方面目前在捷克国家开放数据目录中尚缺失。

5.1 文件格式

在整合多种数据源之前，数据格式是需要考虑的关键方面之一（Abedjan，2018）。不同的数据格式需要采用不同的数据处理方式。此外，每种数据格式对

数据结构也有不同要求，这在原则上会降低数据互操作性。

对于许多常见格式，数据处理可以自动化。然而，如果同一数据表示使用了多种数据格式，处理过程就会变得更加复杂。如果特定文件格式的性质导致数据不一致，处理过程甚至可能需要人工协助，尤其是在还需要更改数据结构的情况下（Heer 等，2019）。

形式化要求 根据 OFS 规范，开放数据集需要以 JSONˇ 格式发布（MVCR，2020）。然而，NODC 上发布的捷克语数据集也采用了其他格式，例如 CSV、Excel 工作簿文件、XML，或 JSON 系列中的特殊格式，这使得数据难以轻松互操作。

分数计算 文件格式 维度根据数据集的数据格式为其评分，得分依据表 1 所示的转换规则确定。拟议表格中的各个等级是根据其格式与所要求的 JSON 格式的相似程度设计的。这些等级也可以理解为表示将特定文件格式转换为 JSON 的难易程度。

表 1：根据数据集的文件格式授予得分的转换表

得分	文件格式
100	JSON, JSON-LD ，GEOJSON
75	XML, GML, KML, RDF
50	CSV
25	XLS, XLSX
0	PDF, TXT

OFS 定义的 JSON 系列文件格式得分最高。XML 系列文件格式的数据集得分次之。XML 文档与 JSON 一样支持层次结构，而且在大多数情况下，可以借助各种在线工具轻松将这种数据格式转换为 JSON。采用表格结构的数据集（例如 CSV 文件）得 50 分。这种文件格式本身不支持层次结构，因此不适用于复杂数据。得分倒数第二低的是采用表格结构、以专有文件格式（如 Excel 电子表格文件）发布的数据集；读取这类数据通常需要专用软件，这与开放数据的普遍理念相悖。对于以无法保证任何结构的格式（例如 TXT 或 PDF）发布的文件，不予评分。

5.2 模式准确性

模式准确性指特征名称相对于 OFS 所定义命名规范的语法准确性。关注建模实体的命名，是侧重互操作性的质量评估中的重要环节：如果两个实体名称不同，即使它们在语义上代表同一对象，也无法自动合并。特征命名错误的一个实际例子是使用英语特征名称位置，而不是捷克语词语 *um´ıstˇen´ı*。

形式化要求 OFS 提出的要求规定了必须包含的必需特征和可选特征的数量，并要求这些特征使用正确的名称。除 JSON-LD 格式规定的唯一键 *@context*（Sporny 等，2020）外，所有特征名称均为捷克语。该键用于将数据集与相应标准互联，以此判断数据集结构是否有效。每个特征名称都只有一种正确形式。

分数计算 数据集的模式准确性得分根据各个特征名称的正确性计算。必需特征和可选特征名称的正确性会以不同方式影响得分，如下方公式 4 所示。

$$score = \sum_{f=1}^{F} w_f n_f \begin{cases} n = 1, & \text{if } n \in N \\ n = 0, & \text{otherwise} \end{cases} \tag{4}$$

其中，f 表示标准规定的特征，w_f表示特征的权重，n_f表示特征名称。OFS 规定的必需特征和可选特征集合记为F，而N 表示受检数据集中所有特征名称的集合。

5.3 模式完整性

模式完整性 专门用于检查各个特征所承载信息的语义正确性。换言之，该维度关注特征的内容，而不考虑其命名是否正确。在这种情况下，也不考虑所使用的数据类型。在实际进行分数计算之前，需要手动调整每个受检数据集——如果特征所承载的信息在语义上符合标准，则必须将原始特征名称修改为与 OFS 规定的词语一致。

形式化要求 该标准定义了一组必选特征和可选特征，这些特征被理解为数据集应包含的信息组成部分，具体内容见第4.2节。这些特征代表了向用户提供关于所建模领域最准确描述所必需的信息。此外，还必须提供与所使用标准相关的信息。

分数计算 模式完整性的得分确定方式与模式准确性维度类似。不过，该维度的侧重点有所不同。最终得分取决于数据集中是否包含预期信息，如公式5所示。

$$score = \sum_{f=1}^{F} w_f i_f \begin{cases} i = 1, & \text{if } i \in I \\ i = 0, & \text{otherwise} \end{cases} \quad (5)$$

其中，f 表示标准指定的特征，w_f表示特征的权重，而i_f表示特征所承载的信息。OFS 指定的一组必选和可选特征记为F，而I 表示根据标准，数据集中预期包含的信息集合。

5.4 数据类型一致性

输入的质量决定了融合数字数据并生成相关信息的能力（Garcia 等人，2018）。由于同类数据可以采用多种格式生成，而且对同一数据使用不同的数据类型会导致耗时的数据清洗，因此，监测数据一致性这一方面是合适的。具体而言，该维度关注同一特征中用于表示相同数据的数据类型是否一致。

所提出评估模型的实现可区分五种主要数据类型，即整数、浮点数、布尔值、字符串和空值。此外，我们决定扩展类型识别范围，增加五种自定义数据类型；这些类型是根据对受评数据集中实际值的分析选出的。自定义数据类型列表如下：

- 统一资源定位符：通过 Python 库*validators*中用于识别统一资源定位符的函数进行识别，
- 电子邮件：通过 Python 库*validators*中用于识别电子邮件的函数进行识别，
- 地址：通过搜索符合捷克格式、由街道名称和号码按给定顺序组成的地址的正则表达式字符串来识别，
- 点：通过一种模式识别，其中地理坐标由点 标签括起，
- 电话号码：通过用于搜索电话号码的正则表达式字符串识别，电话号码可采用捷克语或国际格式。

形式化要求 同一特征中的所有值都应采用相同的数据类型表示。

分数计算 得分受同一特征中所用数据类型的数量及其权重影响。由于数据类型一致性维度关注的是值本身，而非将其与标准的架构进行比较，因此需要对第4.2节所述定义中的必需特征和可选特征作出小幅调整：

- 必选特征：由标准定义为必选，且同时包含在受检数据集中的特征。
- 可选特征：受检数据集中包含的所有其他特征。

换言之，特征权重取决于数据集中所提供特征的总数，而不只是 OFS 定义的特征。假设某个数据集包含标准定义的七个必选特征中的两个，且不包含任何可选特征。那么每个特征的权重为 50。缺失的必选特征不会影响该维度的得分计算。

这意味着，即使数据集缺少某些必选特征，且不包含任何可选特征，在数据一致性方面仍然可以获得最高 100 分。当然，在这种情况下，每个特征中的所有值都必须采用相同的数据类型表示，否则得分会降低。

在单个特征中使用多种数据类型会受到严厉惩罚。得分计算公式见公式 6。

$$score = \sum_{f=1}^{F} w_f \frac{1}{t_f} \quad (6)$$

其中，f 表示数据集中提供的特征，w_f表示特征权重，t_f表示某个特征中使用的数据类型数量。数据集中提供的所有特征集合用F表示。

5.5 数据完整性

我们收集的数据量正在不断增长。然而，并非所有数据都是完整的，其中一些信息可能会缺失。缺失值通常

用空值表示，即用特定标记表明某个值缺失或未定义（Codd，1986），但空字符串也被广泛使用。正如相关文献所述，提供不完整数据主要有两个原因（Codd，1986）。信息对于数据源而言可能是未知的，或者它指的是与特定对象无关的属性。无论是哪种情况，空值都会造成严重的质量问题。它们会给后续数据处理带来困惑，可能导致错误的数据解释，或完全限制数据的可用性。因此，数据处理者需要了解这些空值，以便选择正确的策略对其进行妥善处理。

形式化要求 特征中的所有值理想情况下都应完整；这与空值和空对象不同。

分数计算 数据完整性维度用于衡量非空值的加权比例。这里的比例指所考察特征中的非空值数量除以该特征提供的所有值数量。在所提出的模型中，数据类型为 null 的值和空字符串均视为空值。

至于权重，其计算过程与第 5.4 节所述相同。数据完整性维度的得分按公式 7 所述进行计算。

$$score = \sum_{f=1}^{F} w_f \frac{n_f}{v_f} \tag{7}$$

其中，w_f是特征的权重，n_f表示特征中的非空值数量，而v_f表示所考察特征中的所有值的数量。数据集中提供的一组必需特征和可选特征标记为F。

6 评估

我们应用第 5 节介绍的评估框架，测量旅游兴趣点领域捷克开放数据集的数据质量，重点关注其互操作性以及是否遵循捷克内政部制定的标准。目前，NODC 中发布的数据集仅在元数据层级接受数据质量监测，因此其数据的互操作性质量仍然未知。

总体而言，我们从六个不同的市镇收集了14个数据集，这些数据集均发布在NODC上。由于同一提供方的数据集质量基本相同，我们决定只分析每个提供方的一个数据集，以确保结果不受偏倚。因此，本次评估涉及来自六个不同市镇的六个数据集，详见表2。

每个选定数据集的记录和特征数量各不相同。概览见表3。Huntˇıˇrov市镇的数据集包含10条记录，而布尔诺则提供了内容更丰富的旅游目的地集合，共有350个。就特征而言，布拉格和布尔诺的数据集都记录了每条记录的9项信息。Dˇeˇcˊın的数据集则包含最多的特征，共71项。不过需要注意的是，该数据集中的许多特征都没有有意义的值，只有空值。

表2：提供旅游兴趣点 相关数据的市镇列表及用于分析的选定数据集

市镇	数量 相关 数据集	已选数据集
Brno	2	旅游ˊ地ˊ点
Dˇeˇcˊın	1	兴趣点列表 zˊajmˇu（兴趣点）
赫拉德茨-克拉洛韦	8	Zˊamky
亨特日若夫	1	Turistickˊe cˇle
俄斯特拉发	1	旅游ˊ目的ˊ地
布拉格	1	Vˇyznamnˊe 前瞻性评估点

表3：选定数据集提供的记录和特征数量概览

市镇	数量 记录数	数量 特征
Brno	350	15
代钦ˊ因	254	71
赫拉德茨·克拉洛韦	33	9
洪蒂ˊ罗夫	10	18
俄斯特拉发	66	18
布拉格	323	9

6.1 结果

我们根据提出的五个数据质量方面对每个数据集进行了分析。在每个维度中，数据集的得分范围为0至最高100点。表4列出了各数据质量维度的完整得分记录。图2中的雷达图则直观地展示了结果全貌。

6.2 结果分析

研究结果多种多样；虽然数据集通常在某些维

表4：各市镇在五个监测维度中获得的得分完整结果

市镇	File 格式	模式 准确性	模式 完整性	数据类型 一致性	Data 完整性
Brno	100	0	65.38	53.49	64.46
杰钦	25	0	53.85	58.03	75.99
赫拉德茨-克拉洛韦	100	0.96	38.46	97.94	100
Huntı˘rov	50	0	52.88	70.97	76.3
俄斯特拉发	50	0	52.88	65.73	81.03
布拉格	100	0	25	78.56	100

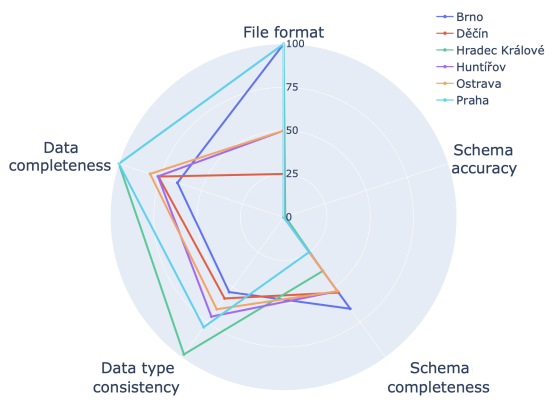


图2：各市镇在每个数据质量维度中所取得得分的比较

度（例如数据完整性）上取得了较为令人满意的得分，但在其他方面，尤其是模式准确性方面，所有数据集都表现出显著不足。在接下来的段落中，我们将重点分析所识别出的缺陷最多的维度，并对结果进行深入分析。

模式准确性 所有数据集在模式准确性 维度上均未取得令人满意的结果，原因有很多。经过对特征名称的全面审查，我们确定了以下原因：

- 两个数据集（布尔诺、赫拉德茨·克拉洛韦）使用英语特征名称，
- 两个数据集（Huntı˘rov、俄斯特拉发）使用捷克语特征名称，但未使用变音符号，也未以其他方式进行修改，
- 两个数据集（D˘e˘cın、布拉格）混合使用英语和捷克语特征名称，即使未使用变音符号，也未以其他方式进行修改。

我们观察到，每个市镇都使用自己独有的一组特征名称。然而，在数据集中仍可看到某些相似之处。例如，俄斯特拉发和Huntı˘rov的数据集共享18个特征名称中的10个，尽管这些名称与标准所要求的特征名称不同。

模式完整性结果表明，尽管特征命名不准确，受检对象数据集至少在一定程度上包含标准所要求的信息。

我们认识到，数据集通常以地点列表的形式提供，且仅关注其位置，这是造成模式完整性方面上述困难的主要原因。对于以GEOJSON格式发布、专门为此目的设计的数据集而言，情况尤其如此。但即使某个数据集对于一种用途来说质量足够，也未必适用于另一种用途（Sadiq 和 Indulska，2017）。因此，最好补充缺失信息，以增加在旅游业中使用这些数据的可能性。

7 结论

本文提出了一种面向互操作性的质量评估框架，包含五个数据质量维度。数据质量维度的选择基于数据集的互操作性以及对指定标准（即 OFS）的遵循情况。为评估该框架的适用性，我们评估了旅游兴趣点数据集的质量。这些数据集从捷克开放数据国家目录下载，评估结果揭示了该目录中的质量问题。

尽管受评估的数据集在数据完整性和数据类型一致性 维度上仅存在中等程度的不足，但模式准确性 的结果却是所有维度中最差的。所有数据集在这一维度上的表现都很差，这表明选定数据集在特征命名规范方面未遵循标准。此外，大多数数据集仅将单个项目建模为普通地点，而非要求的旅游对象。因此，捷克开放数据集目前在旅游语境中实际上无法实现互操作性，其价值共创潜力也因此降低。

尽管本文主要关注开放

数据质量评估在捷克共和国的实践，但研究发现对于任何旨在改进开放数据集质量评估流程的国家都具有参考价值，因为设计过程中借鉴不同国家的经验至关重要。一旦我们了解开放数据集的质量并识别其质量缺陷，就能指导数据生产者提供并改进具有影响力的数据，使社会能够充分利用具备互操作性的开放数据。

致谢 本研究由欧洲区域发展基金“网络安全、网络犯罪与关键信息基础设施卓越中心”（编号：CZ.02.1.01/0.0/0.0/16 019/0000822）资助。

参考文献

Abedjan, Z. (2018). 大数据技术百科全书，章节“数据分析”，第 1–6 页。施普林格国际出版。 Berners-Lee, T. (2012). 五星开放数据。 https://5stardata.info/en/ 访问日期：2022-03-24。 Bizer, C., Heath, T., and Berners-Lee, T. (2011). 语义服务、互操作性与 *Web* 应用程序：新兴概念，章节“关联数据：迄今为止的历程”，第 205–227 页。IGI Global，Hershey, PA。Carrara, W., Enzerink, E., Oudkerk, F., and Radu, C. (2018). 数据管理者与数据持有者开放数据指南。欧盟出版局。Codd, E. F. (1986). 关系数据库中的缺失信息（适用与不适用）。 *SIGMOD Rec.*，15(4):53。Dvoˇr´ak, M., Sp´al, M., Marek, J., and Kl´ımek, J. (2020). Otevˇren´e form´aln´ı normy (OFN) – doporuˇcen´ı 1. ˇcervence 2020。捷克共和国内政部。Garcia, J., Molina, J. M., Berlanga, A., and Patricio, M. A. (2018). 大数据技术百科全书，章节“数据融合”，第 1–6 页。施普林格国际出版，查姆。Ge, M., Chren, S., Rossi, B., a nd Pitner, T. (2019). 智能电网系统的数据质量管理框架。载于 商业信息系统——第 *22* 届国际会议，*BIS 2019*，西班牙塞维利亚，第 354 卷 *LNBIP*，第 299–310 页。Ge, M. and Lewoniewski, W. (2020). 协作式开放数据质量模型的开发。载于 第 *24* 届 *KES-2020*国际会议论文集，虚拟会议，*2020* 年 *9* 月 *16–18* 日，第 176 卷 *ProcediaComputer Science*，第 1883–1892 页。Elsevier。Heer, J., Hellerstein, J. M., and Kandel, S. (2019). 大数据技术百科全书，章节“数据整理”，第 584–591 页。施普林格国际出版。 Kl´ımek, J. (2019). 捷克国家开放数据目录的 DCAT-AP 表示及其影响。*Journal of Web Semantics*，55:69–85。Kl´ ımek, J. (2020a). Otevˇren´a data a otevˇren´e form´aln´ı normy。捷克共和国内政部。https://data.gov.cz/%C4% 8Dl%C3%A1nky/otev%C5 %99en%C3%A9-form%C3% A1ln%C3%AD- normy-01-%C3%BAvod 访问日期：2022-03-18。Kl´ımek, J. (2020b). Otevˇren´e form´aln´ı normy (OFN)。捷克共和国内政部。 https://data.gov.cz/ofn/ 访问日期：2022-03-18。Kl´ı mek, J. (2022). Pˇr´ıklady ˇspatn´e praxe v oblasti otevˇ ren´ych dat。捷克共和国内政部。 https://opendata.gov.cz/%C5%A1patn%C3%A1- praxe:start 访问日期：2022-03-24。MV ˇCR (2020). 作为提供者，我应如何使用开放形式规范。捷克共和国内政部。 https://data.gov.cz/ofn/pou%C5%BEit%C3%AD- poskytovateli/ 访问日期：2022-03-18。Nikiforova, A. (2020). 开放数据质量。*CoRR*，abs/2007.06540。OHA (2021). 公共数据基金。捷克共和国内政部首席电子政务架构师办公室。https://archi.gov.cz/en:nap:verejny datovy fond。访问日期：2022-03-18。Sadiq, S. and Indulska, M. (2017). 开放数据：质量重于数量。*International Journal of Information Management*，37(3):150–154。 Sporny, M., Longley, D., Kellogg, G., Lanthaler, M., Champin, P., and Lindstr¨om, N. (2020). JSON-LD 1.1——用于关联数据的基于 JSON 的序列化。万维网联盟。 https://www.w3.org/TR/json- ld11/#the-context。访问日期：2022-03-18。Thereaux, O. (2020). 数据与 COVID-19：为何标准至关重要。开放数据研究所。 https://theodi.org/article/data-and-covid-19-why- standards-matter/ 访问日期：2022-03-18。Ubaldi, B. (2013). 开放政府数据。（22）。Vetr`o, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., and Morando, F. (2016). 开放数据质量测量框架：定义及其在开放政府数据中的应用。*Government Information Quarterly*，33(2):325–337。Viscusi, G., Spahiu, B., Maurino, A., and Batini, C. (2014). 遵守开放政府数据政策：对意大利地方公共行政机构的实证评估。V´ıta, M. (2021). Vybran´e techniky anal´yzy datov´e kvality a jejich aplikace na data z centr´aln´ı evidence projekt˚u vavai。捷克共和国内政部。https://data.gov.cz/%C4%8Dl%C3% A1nky/techniky- anal%C3%BDzy-datov%C3% A9-kvality-cep 访问日期：2022-03-18。Wang, R. Y. and Strong, D. M. (1996). 超越准确性：数据质量对数据消费者意味着什么。*Journal ofManagementInformationSystems*，12:5–33。