MASARYK UNIVERSITY

FACULTY OF INFORMATICS

Analysis of Open Data's Standardisation

Master's Thesis

BC. DAŠA KUŠNIRÁKOVÁ

Brno, Fall 2021

MASARYK UNIVERSITY

FACULTY OF INFORMATICS

Analysis of Open Data's Standardisation

Master's Thesis

BC. DAŠA KUŠNIRÁKOVÁ

Advisor: Ing. Leonard Walletzký, PhD.

Department of Computer Systems and Communications

Brno, Fall 2021

FACULTAS ARTIS INFORMATICÆ
UNIVERSITAS MASARYKIANA

Declaration

Hereby I declare that this paper is my original authorial work, which I have worked out on my own. All sources, references, and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Bc. Daša Kušniráková

Advisor: Ing. Leonard Walletzký, PhD.

Acknowledgements

I would like to thank my advisor Ing. Leonard Walletzký, Ph.D., for his guidance and for giving me the opportunity to explore the topic of open data in depth. Special thanks go to doc. Mouzhi Ge, Ph.D., for all his insightful advice.

I would also like to thank all Laboratory of Service Systems members for creating a friendly environment and effective collaboration. My gratitude goes especially to Bc. Dominik Skála for his assistance in finding datasets and daily reminders of how many days remain until the deadline.

Finally, I am thankful to my family and friends for their support during my whole studies.

Abstract

This thesis aims to analyze the actual situation regarding open data standardization, whose primary goal is to improve its quality and ensure data interoperability. The theoretical part introduces the currently-used data quality assessment methods, followed by the existing standardization solutions within Europe. The main focus is on the tools developed by the Ministry of Interior of the Czech Republic. The practical part of this thesis then examines the so far unknown cell-level quality of available datasets on a selected topic, and the actual adherence to the developed standards. The quality of individual datasets is determined in a total of five different dimensions by a proposed evaluation model.

Keywords

open data, smart city, data quality, data interoperability, standardization, open standards

Contents

Introduction 1

1 Smart city concept 3

2 Open data 4
- 2.1 Definition . . . 4
- 2.2 Benefits . . . 5
  - 2.2.1 Economy . . . 6
  - 2.2.2 Social welfare . . . 6
  - 2.2.3 Performance . . . 7
- 2.3 Current open data situation . . . 7
  - 2.3.1 European countries in general . . . 7
  - 2.3.2 Czech Republic . . . 8
- 2.4 Barriers . . . 8
  - 2.4.1 Technical barriers . . . 9
  - 2.4.2 Awareness barriers . . . 10
  - 2.4.3 Practical barriers . . . 10

3 Data quality 12
- 3.1 Definition . . . 12
- 3.2 Quality assessment tools . . . 13
- 3.3 Open data quality assessment in the Czech Republic . . . 16
  - 3.3.1 Current state . . . 16
  - 3.3.2 Future plans . . . 18

4 Standardization 19
- 4.1 Ways of standardization . . . 19
  - 4.1.1 Metadata standards . . . 19
  - 4.1.2 File format standards . . . 20
  - 4.1.3 Schema standards . . . 21
- 4.2 Open formal standards . . . 22
  - 4.2.1 Development process . . . 22
  - 4.2.2 OFN features . . . 23
  - 4.2.3 Current state of OFN adoption . . . 25
  - 4.2.4 Motivation techniques . . . 26

4.3 Summary 28

5 Model proposition **29**
5.1 Research approach 29
5.2 Terms definition 30
5.3 Data quality dimensions 32
5.3.1 File format 32
5.3.2 Schema accuracy 36
5.3.3 Schema completeness 37
5.3.4 Data type consistency 39
5.3.5 Data completeness 41

6 Evaluation **42**
6.1 Datasets 42
6.2 Evaluation results 44
6.3 Results analysis 46
6.3.1 File format 46
6.3.2 Schema accuracy 48
6.3.3 Schema completeness 49
6.3.4 Data type consistency 50
6.3.5 Data completeness 51
6.4 Summary 52

7 Conclusion **54**
7.1 Future work 55

Bibliography **56**

A Complete JSON schema **62**

B Recommendations for examined datasets **65**

List of Tables

5.1 List of mandatory and optional features and their weights for a dataset regarding *Tourist point of interest* specified by OFN 31
5.2 Conversion table for awarding scores based on file format 33
6.1 List of municipalities providing data regarding *Tourist points of interest* and selected datasets for analysis 43
6.2 Overview of the number of records and features provided by the selected datasets 44
6.3 Complete results of earned scores in five monitored dimensions for individual municipalities 44

List of Figures

2.1 Main barriers for the re-use of open data identified by public sector representatives [10] 9

3.1 Visualization of the 5-Star Rating System. Taken from `https://5stardata.info/en/`. 14

4.1 An example of OFN scheme for objects regarding *Tourist points of interest*. Translated to English [37]. 24

4.2 A snippet of the JSON schema for the sample OFN *Tourist points of interest*. 25

4.3 Example of a leaflet (left) generated by the demo application according to the given data (right) 27

5.1 Example of a semi-structured XML element 35

5.2 Sample data about a tourist destination in its original form. The information about the object's *name*, *description*, *location* as well as the *identifier* are contained in the dataset but their naming is incorrect. This causes that the presence or absence of such features cannot be inspected automatically. 38

5.3 Features' names which semantically matched the information specified by the standard were renamed. Features *název*, *popis*, *umístění*, and *iri* now correspond with the standard, and therefore, the presence of expected features can be checked against the OFN. However, the dataset lacks data about the opening hours or entrance fee, which lessens the dataset's score in terms of schema completeness. 38

6.1 Radar charts visualizing the achieved scores in each data quality dimension for individual municipalities 45

6.2 Comparison of the achieved scores in each data quality dimension for individual municipalities 46

6.3 Statistical overview of the achieved scores aggregated by the proposed data quality dimensions 47

6.4 Overview of the most commonly present and absent information carried by mandatory features in the examined datasets 50

6.5 Bar chart visualizing the number of used data types by mandatory and optional features 51

6.6 Bar chart visualizing the ratio of non-null values within mandatory and optional features 52

Introduction

Since the publication of the Directive 2003/98/EC on the re-use of public sector information, the European Union has encouraged member states to make as much data stored by public sector organizations available for re-use as feasible. Such opening data to the public has the potential to bring enormous benefits to citizens, public administrations and businesses while also facilitating better cooperation throughout Europe. The main advantages of open data publication include not only improving the quality and transparency of public services and their delivery. Open data can also significantly impact economic growth and society, because it contributes to the development of novel services and increases citizens' participation in policymaking [1].

With an increasing number of published datasets, relying just on the amount of data is no longer sufficient, and the importance of data quality is growing. Low data quality represents a major issue for interoperation and hence making use of the potential of open data. However, the data quality can be viewed from different perspectives in the open data environment, resulting in various quality requirements and assessment methods used by European open data initiatives.

In general, accessibility and representational dimensions, such as availability, licensing, or data format, are put under examination almost constantly, but cell-level quality checks are relatively scarce. These could significantly increase the quality of open data, especially in terms of data completeness, semantic accuracy, or value and schema consistency. All of these are the main driving forces behind the development of open standards.

Open data standards are reusable agreements created and maintained through a collaborative decision-making process. They aim to improve the data quality by establishing rules on data sharing, specifying common formats and vocabulary and describing how data should be structured. There are several open data standard directories across Europe. However, their adoption is insufficient. Only a few data providers follow these standards, which has resulted in a state where some of them are no longer maintained. Consequently, data

sharing issues still persist, and the open data's full potential remains unrealized.

Currently, open data standards are being introduced in the Czech Republic, too. With their focus on the dataset's content as well as semantics, they aim to increase the quality of Czech open data and ensure its interoperability across datasets from various data providers. However, the actual adoption of these standards is questionable, although they are legally binding on data providers. What is more, even the general cell-level quality of Czech open datasets is unknown. Only the quality of their metadata is monitored in the Open Data National Catalogue at this time, leaving the content and structure to the sole responsibility of the data providers.

Therefore, this thesis aims to analyze the current situation regarding open data standardization within Europe while focusing mainly on the solutions introduced in the Czech Republic. The next goal is to provide a fundamental view on the so-far unknown cell-level data quality of Czech datasets on a selected topic which are published in the Open Data National Catalogue.

The text of this Master's thesis is divided into seven chapters. The first chapter sets the work into the smart city context. The second chapter overviews the domain of open data, its advantages, and barriers to its wider usage. The third chapter is dedicated to the concept of data quality and the European assessment methods, which are currently used for data evaluation. The Czech approach to the data quality assessment is also presented. The fourth chapter discusses the motivation behind data standardization and the different levels at which standardization can operate. A large part of that chapter is devoted to *Open formal standards*, a standardization solution for open data introduced by the Czech Ministry of the Interior. The fifth chapter introduces the proposed fundamental model for the open data quality assessment in terms of their interoperability and adherence to the established standards. The evaluation results are then presented in the sixth chapter. The final chapter summarizes the findings of the work and offers suggestions for further research.

1 Smart city concept

According to the United Nations[1], urban areas house 55% of the world's population, and the number is expected to increase to 68% by 2050. With such a large inflow of people, society will need to develop innovative methods to effectively manage the new living conditions. As the world becomes more urban, cities must become smarter.

The concept of smart cities became popular in the early 2010s to highlight how technology combined with data can improve urban planning and city management [2]. A smart city is in the literature defined as a city that effectively mobilizes and utilizes existing resources, such as finances, natural resources, social and cultural capital, information, and technology, to achieve its goals [3]. These goals cover multiple areas, including:

- **People**: enhancing the life quality for its residents, commuting workers and visitors,
- **Environment**: improving resource efficiency, reducing environmental pressure, and increasing resiliency,
- **Prosperity**: fostering an innovation-driven, green economy,
- **Governance**: supporting the development of local democracy.

Today, more data is being accessible than ever before. And as data fosters knowledge and innovation, it is the driving force also behind the development of smart cities. It is, however, essential to learn how to use the data effectively and take what is relevant [4].

Using data in smart cities entails not only its producing or collecting, but also its sharing among individual people, private companies, as well as with the government and general public [5]. As a result, new challenges that municipalities need to deal with have arisen. In particular, the ability of data to interoperate smoothly, and the data quality in general, has started to be put in question. The potential of inconsistent or low-quality data to provide any value is limited.

---

1. `https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html`

2 Open data

In the previous decade, there has been a growing global trend for systematic data openness. The first open data government policy was launched in 2009 in the United States [6]. Since then, national, regional, and local governments worldwide have continued to embrace open data policies and plans and launch open data projects.

2.1 Definition

The Open definition created by the Open Knowledge Foundation[1] defines open data as *"data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and share-alike"* [7]. The following paragraphs discuss the definition in a more detailed manner.

On the one hand, the meaning of the term "open" itself is explicit. It unambiguously covers two dimensions of data openness [8]:

- **Legal**: Data must be published under an open licence with minimal restrictions for re-use.
- **Technical**: Data must be published in an electronic, machine-readable, and non-proprietary file format to enable access and use of the content without restrictions to a non-open source software tool, where possible.

However, the fact that open data is freely available does not necessarily mean the data is available free of charge [8]. The current European Directive 2019/1024 on open data and the re-use of public sector information states that the zero-cost method is recommended but not strictly necessary. Where the fee is needed, it should be limited only to the marginal costs of data reproduction, provision, and dissemination.

1. the world's largest open data community (`https://okfn.org`)

On the other hand, it is essential to note that the open data movement establishes a set of norms and principles that may be applied in various ways depending on the country and the legal framework in place [8].

As for the Czech Republic, the meaning of open data is anchored in law in Act No. 106/1999 Coll. on Free Access to Information as *"information published in a manner enabling remote access in an open and machine-readable format where the manner and purpose of the subsequent use of such information is not restricted, and the information is recorded in the Open Data National Catalogue"*.

Based on the information presented above, multiple key characteristics for Czech open data can be summarized. Data can be considered open if it is:

- publicly available on the Internet,
- freely accessible and can be used or redistributed by anyone and for any purpose,
- published under an open licence explaining the conditions under which the data may be used,
- recorded in the Open Data National Catalogue[2],
- published in a machine-readable format[3],
- provided with complete documentation,
- provided with curator's contact information,
- its content is accessible with an open-source software tool.

2.2 Benefits

The motivation behind the publication of open data is that the data provides significant benefits to citizens, businesses, and public authorities, as well as facilitating closer cooperation throughout Europe [9].

2. `https://data.gov.cz/datové-sady`
3. e.g. CSV, JSON, XML, RDF; machine-readable formats do not include PDF

The impact of open data can be assessed from different perspectives. These can be divided into three main categories, in particular *economy*, *social welfare* and *performance*. A short description of each of these aspects is provided in the following sections of this chapter.

2.2.1 Economy

From the economic point of view, open data can provide both direct and indirect benefits. Direct benefits can be understood as monetized advantages realized in market transactions. They can be in the form of revenues and gross value added, cost savings in public administrations or the number of new work opportunities involved in the production of a service or product. Indirect economic advantages include creating new goods and services, greater efficiency in public services, and growth in related markets [10].

It is challenging to accurately quantify the economic advantages of open data as the essential benefits are indirect. However, a study prepared by the European Data Portal expects the European open data market size will have increased from €184 billion (2019) up to €334 billion by 2025 in case of an optimistic growth scenario. As a result, an annual growth rate up to 10.4% for the number of jobs generated is expected [11].

2.2.2 Social welfare

Regarding the social point of view, there are various benefits to be mentioned. For example, open data raises citizen participation in policy-making. Individuals can analyze the data behind published reports, and based on this, make better-informed decisions. What is more, open data even allows citizens to participate in creating policies that are designed to their needs, or aimed for marginalized groups to integrate into society.

Open data also creates the potential for closer collaboration between the private and the public sector. That can be realized, for instance, by organizing hackathons, where interested parties solve specific issues based on open data [1].

2.2.3 Performance

Open data has the potential to improve also the efficiency of public services. One of the main benefits in terms of performance is spending transparency, resulting in better resource allocation and allowing companies to concentrate resources on their core business activities.

Besides that, increased efficiency can be achieved by faster access to information in general [9]. Because the data is freely available, locating, getting, and re-using data takes less time.

2.3 Current open data situation

The advantages of opening data are undeniable. However, it is somewhat challenging to get to them in practice, as data opening is a long process requiring thorough preparation. This section overviews the current open data situation within Europe as well as the Czech Republic in particular.

2.3.1 European countries in general

According to the most recent Open data maturity report 2020, which measures the degree of openness of individual European countries, a total of seven countries achieved excellent results overall, such as Denmark, Estonia, or Poland. All these nations have adopted an advanced open data strategy, with good coordination of open data initiatives at all government levels. Therefore, they fall into the category of trendsetters [12]. These countries are characterized by a national portal providing a variety of features for advanced users and data publishers, a good level of data quality while planning its further improvement, and mechanisms to assess the open data's impact in different domains.

On the contrary, several other European nations rank among open data beginners. Their common features include an open data portal with only limited functionality, and no monitoring of open data reuse nor assessment of its effect in the country. In terms of data quality, these nations take only a few steps to enable higher-quality data to be published. Portugal, Hungary or Slovakia belong to such a group.

2.3.2 Czech Republic

In 2020, the Czech Republic's overall rank has slightly declined compared to other countries and prior years [12]. Although the state excels in implementing data-opening policies, laws, strategies, and action plans [13], the reality and subsequent implementation lag, especially in the area of the data portal and its utilization. In particular, Czechia ranked as the second-worst in Europe in this regard. The implementation problem can also be observed directly from the NKOD; more than 18 000 government institutions are currently registered by the Czech Ministry of Finance[4], but only 45 of them publish data in an open format in the Open Data National Catalogue[5].

The statistics presented above demonstrate that the current open data situation in the Czech Republic is not ideal and therefore offers room for improvement, especially in terms of the number of active data providers. However, the open data potential can only be realized when the data is accessible, re-usable, and transformed into value [14]. In addition to the issues on the provider side, a number of challenges need to be overcome on the data re-user side, too. These barriers are discussed in the following section.

2.4 Barriers

Lack of active data providers in the Czech Republic and thus relatively low number of published datasets in the Open Data National Catalogue is not the only issue data re-users need to face. The following paragraphs aim to introduce other types of problems regarding open data usage, namely technical, practical, and awareness barriers.

4. List of Czech government institutions (October 2021): `https://www.mfcr.cz/cs/verejny-sektor/rozpoctove-ramce-statisticke-informace/verejny-sektor/sektor-vladnich-instituci/2021/seznam-vladnich-instituci-v-cr-rijen-20-43418`
5. List of Open data publishers (October 2021): `https://data.gov.cz/publishers`

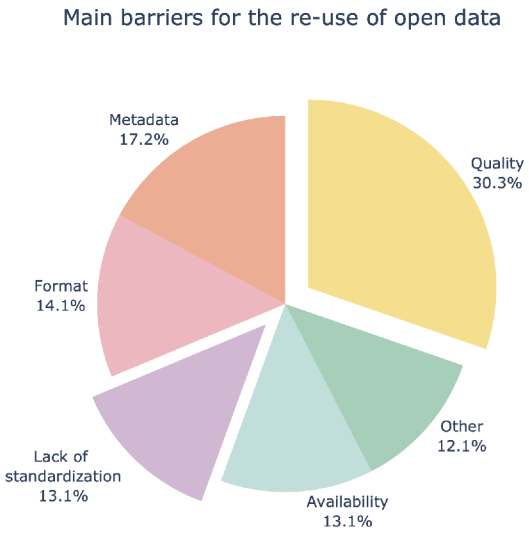

Figure 2.1: Main barriers for the re-use of open data identified by public sector representatives [10]

2.4.1 Technical barriers

The technical viewpoint focuses on the technologies and infrastructures related to the utilization of open data. A survey taken by the European Data Portal as a part of a research project on barriers for the re-use of open data [10] showed that overall 68% of the public sector representatives experience difficulties with data quality, which is followed by data availability (39%) and lack of standardization concerns identified overall by 32% of respondents.

In case we focus only on the most serious issue recognized by public sector representatives, quality or lack of standardization were identified as the main issue by more than 43% of respondents. A visualization of the survey results is presented in Figure 2.1.

Availability Even though the number of published datasets is constantly growing, open data availability remains an issue for Europan public sector representatives. Again, this result is supported by the relatively small number of Czech data providers, as presented in the statistics in section 2.3.

Quality The open data quality issues seem to be even more concerning. This perspective covers especially data provided in various formats and different structures.

The Czech Open Data National Catalogue solves some of the problems covered in the European survey results. For instance, datasets' metadata are checked and regularly analyzed[6], which simplifies quality checks for individual data providers and thus helps increase data quality for data re-users. Other tools are used to ensure open data quality as well, discussed more in detail in section 3.3. However, data quality checks on the datasets' cell level are not performed. Therefore, the data quality barriers as identified by European public sector representatives are still relevant, even for the Czech Republic itself.

2.4.2 Awareness barriers

As shown in section 2.2, the potential of open data is enormous. However, it can only be realized if both data publishers and consumers are convinced of the advantages, but public awareness of open data and its importance remains unsatisfactory [15]. One explanation for the low level of awareness is that open data represent an abstract concept. The benefits of opening data and its possible added value are in general unclear to the general public, and potential re-users [14].

On the other hand, if data providers are more aware of the open data concept and its value, they also become increasingly aware of what can be done with it. Knowing how the released data may be further utilized allows data providers to adjust their publishing strategy to this need, resulting in more quality and more acceptable open data for those who can convert it into value. Hence even data re-users may play an important role in raising open data awareness, in particular by sharing their experience and demonstrating their ways of using open data.

2.4.3 Practical barriers

Even if potential data re-users are aware of the benefits of open data, they often need to deal with practical obstacles. To effectively use the

6. https://opendata.gov.cz/statistika:datova-kvalita

data, users must possess particular hard and soft skills [16]. Working with open data necessitates these particular skills [14]:

- **Analytical**: Required for problem analysis and solution proposal with the help of data.
- **Statistical**: Needed for detailed data analysis.
- **Technical**: Important for actual working with the data, its downloading and processing.
- **Business insights and domain knowledge**: It is necessary to understand the market relevant to the particular business.

However, these abilities are not always fully available, especially in public sector organizations. Re-users usually lack business insights and marketing skills if they are technically skilled [10], or vice versa.

3 Data quality

With the broader adoption of the open data paradigm worldwide and the increasing number of published datasets, the focus has shifted, and their quality has started to be questioned, too [17]. Such uncertainty about data quality is an issue because the potential of low-quality datasets to create any value is poor.

One of the key problems to mention is the interoperability of open datasets, as the concept of open data is based on the needs of reusability and interconnection with other data. Because the data comes from different publishers, its structure or individual values are often incompatible with each other. Such diversity of datasets, e.g., in terms of data formats or vocabularies used, then significantly increases the effort required for further data usage [18], or may even make data merging completely impossible.

Despite interoperability issues, other risks related to the usage of low-quality open data need to be mentioned. Because open data consumers are usually not their producers, there is no well-defined process for data cleansing, resulting in potentially incorrect data curation and its transformation [19]. Consequently, users may expend substantial effort to obtain meaningful results from data only to discover that the findings are insufficient. In the worst case, users may not even recognize that the data is of poor quality and mistakenly rely on inaccurate results [17], and reduce the quality of other outputs that are based on this data.

This chapter reviews the open data quality definition and presents the widespread quality assessment tools in the European Union. In the end of this chapter, the current process of open data validation in Czech Republic regarding data quality is presented.

3.1 Definition

The European Data Portal claims that the quality of open data stems mainly from its usability and that *"Open data becomes usable when a human can understand it and a machine can manipulate it."* [20]. Community-driven standards further elaborate on this idea and define more partic-

ular requirements for an open dataset to be useful. They are devoted to four main areas and aim to ensure these aspects:

- **Technical**: Defines technical requirements, such as the format and structure of the data.
- **Legal**: Sets legal requirements on open data like protecting personal information and other sensitive data, publishing datasets under open license, and preserving data owners' rights.
- **Practical**: Aims to ensure that the open data can be found and relied upon by users; these include regular data updates and commitment to continue making data available.
- **Social**: Aims to ensure data sustainability by having an engaged community of users; these contain having discussion groups and forums and publishing how-to guides on working with the data.

3.2 Quality assessment tools

In order to check the quality of open data, there are various tools to mention. For example, the European Data Portal points to the 5-Star Open Data rating system and Open Data Certificate to assess the quality of open data. Both place different quality requirements on open datasets.

5-Star Open Data Rating System The 5-Star Open Data is a rating system suggested by the W3C Foundation[1] in 2010. The data subjected to quality analysis is awarded a certain number of stars, corresponding to its quality level. The quality requirements for each additional star cover the criteria for all previous stages, and the maximum possible amount of awarded stars for data of the highest quality is five [21]. The individual levels of the 5-Star Rating System are listed below, and a visualization is provided in Figure 3.1.

1. https://5stardata.info/en/

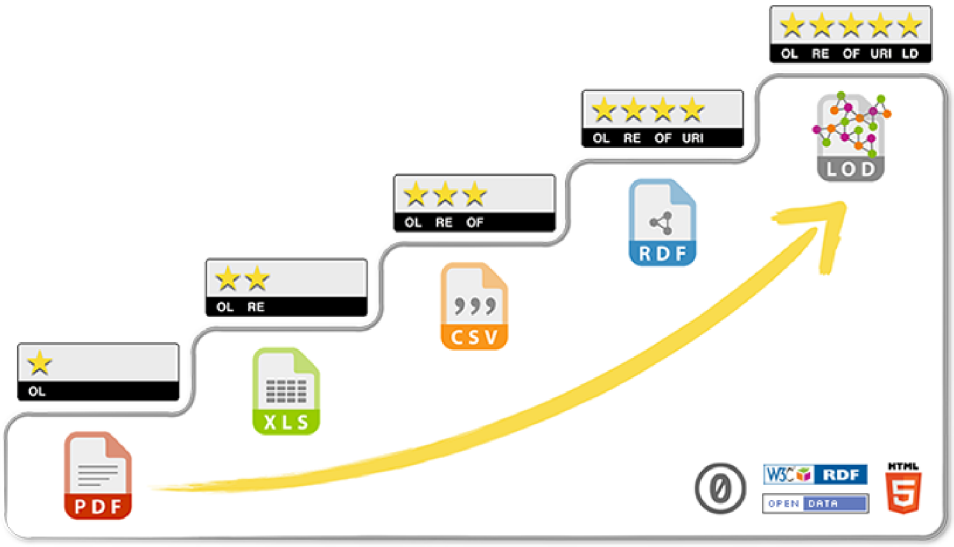

Figure 3.1: Visualization of the 5-Star Rating System. Taken from `https://5stardata.info/en/`.

- **1***: The data is publicly available on the Internet under an open licence and in any format.
- **2***: The data is published as structured data in format that is machine-readable (e.g. a Microsoft Excel file instead of a table provided in a scanned image).
- **3***: The data is published in a non-proprietary format (e.g. a CSV file instead of a Microsoft Excel file).
- **4***: The data items are denoted with URIs so they can be shared.
- **5***: The data contains links to other data to provide context.

Open Data Certificate Open Data Certificate[2] is a self-assessment tool based on a questionnaire developed by the Open Data Institute[3]. It is utilized mainly by British public organizations, such as the UK's Office

2. https://github.com/theodi/open-data-certificate
3. https://theodi.org

for National Statistics and UK Data Service. The tool allows publishers to rate the quality of their data. The application evaluates the datasets using several criteria, including legal and technological considerations. Based on the evaluation results, the datasets are then awarded a bronze, silver, gold, or platinum certificate [22]. The requirements for obtaining a particular certificate type are as follows:

- **Bronze**: Openness, accessibility and reusability are checked.
- **Silver**: Except for the bronze quality requirements, the data is also trustworthy and comes with continuous support from the publisher.
- **Gold**: Besides silver requirements, the data is documented and regularly updated; it also offers better support level and contains machine-readable legal statements.
- **Platinum**: On top of the gold-level characteristics, platinum level checks for unique IDs, provenance documentation and a dedicated support team.

Even though the two tools presented above represent the main tools used for data quality assessment in Europe, their results may be misleading as the evaluation focuses only on a small subset of data quality dimensions. The evaluation through the simplicity of use and openness, as done by the 5-Star Open Data Rating System, may not reflect all user demands on data quality [23].

As stated before, the quality of open data is currently evaluated mainly by the base technical and legal standards on which the data is published; that is mostly on the metadata level. However, in order to increase the interoperability of data and thus increase the potential of creating value, the focus must shift to the data itself. Data quality dimensions such as data completeness, semantic accuracy, schema consistency, or timeliness should be therefore also considered when assessing a dataset's quality.

3.3 Open data quality assessment in the Czech Republic

As mentioned in chapter 2, the Czech open data definition specifies that all open datasets must be registered in the National Open Data Catalogue (NKOD). The primary aim of NKOD is to collect published data and information from the entire public administration in one place and thus ensure their maximum availability for further use. Because the data is present in one place, gaining a comprehensive view of this data is possible.

The following section utilizes this view and presents the current state of open data in the Czech Republic with respect to the assessment of data quality.

3.3.1 Current state

The National Open Data Catalogue does not check individual datasets' quality; NKOD only works with the metadata of the corresponding data files. Even though the Ministry of the Interior supports data providers in improving the quality of their data through various tools, cell-level data quality of individual datasets is solely the responsibility of data providers [24]. They must pay attention to the quality of their data, check it on a regular basis, and take action to resolve any discovered shortcomings. Mentioned tools intended for helping improve open data quality, which serve as guidelines for data providers, are listed below.

Metadata quality The Ministry of the Interior fully utilizes the benefit of having all open datasets published in one location. The availability of resources registered in the NKOD[4], as well as statistics on the quality of their metadata information[5], are measured on a daily basis. The results concerning the metadata quality or availability of datasets from various data providers are then published on dedicated web pages, so their current state can be easily monitored.

The current state of metadata quality of datasets published in NKOD was presented at the conference *Open Data 2021*[6], organized

4. `https://opendata.gov.cz/statistika:dostupnost`
5. `https://opendata.gov.cz/statistika:datova-kvalita`
6. `https://opendata.gov.cz/edu:konference:2021`

by the Czech Ministry of the Interior. The results showed that only 33% of all published datasets have their metadata filled out correctly, while the others report serious flaws [25]. For example, 78% of data providers publish datasets without filling in all required metadata, and 44% of data providers have some of their published datasets unavailable.

Bad practice examples The Czech Ministry of the Interior gathers examples of poor data opening practices and provides methods for their correction. The most common mistakes are thus listed in one location to help others avoid them. They are divided into five categories, each of which covers a particular area, in particular [26]:

- **Legal barriers**: Cover bad practice examples regarding licensing and liability disclaimer.
- **Documentation and metadata issues**: Deal with issues with documentation or metadata, such as mismatched keywords.
- **Technical barriers**: Include solutions for problems with unsecured protocols, wrong HTTP headers, or the usage of machine unreadable data formats.
- **Data formats**: Cover the most frequent issues when publishing CSV and XML data.
- **Data content issues**: Concern issues regarding the content of data, such as the wrong usage of identifiers or missing coordinate reference system.

From all the five categories listed above, only the last one deals with the cell-level of published datasets; the other ones treat rather technical and legal aspects regardless of the content of the datasets. Even though the data content is partially covered in the *Data content issues* category, checks solely of the coordinate system or identifiers used in the datasets cannot ensure full interoperability between data published by different data providers. This is why other methods are needed to assess the open data quality on the cell level to a much more considerable extent. Open formal standards attempt to approach such a state.

Open formal standards Open formal standards are agreements whose observance helps improve open data interoperability in the most significant measure compared to the previously mentioned tools.

Because OFNs represent the primary focus of this work, this tool is presented in greater detail in section 4.2.

3.3.2 Future plans

In addition to the presentation on the current state of open data in the Czech Republic, future ambitions were also discussed at the conference *Open Data 2021*. Because the current state of the dataset's metadata quality is unsatisfactory, the ministry plans to focus on increasing the quality of published datasets and, in general, sharing awareness of the benefits of data opening.

This is to be achieved by organizing workshops or providing individual consultations [25]. Data providers and the general public interested in open data are also offered to complete e-learning courses[7], which have been launched recently.

Other plans concern the technical side of the NKOD, where automatic checks of adherence to OFNs would be enabled while publishing datasets. Such an improvement might significantly improve the dataset's quality on the cell level and thus guarantee the interoperability of datasets from different providers. However, no specific timeframe for this adjustment has been established yet.

7. `https://data.gov.cz/vzděláván%C3%AD/`

4 Standardization

Open data has become a standard technique for many public entities to share information with the public. With the expanding number of open data providers and the diversity of data available in this manner, it appears that data unification is an essential precondition for data usability.

The general public recognizes the need for data standardization, too. Based on the results from the discussion on the European strategy for data held in July 2020 [27], almost 80% of the 512 respondents declared that utilizing data from other data providers was problematic. Besides that, over 91% of respondents agree that standardization is necessary to facilitate interoperability, while 50% believe open standards should be funded by national government bodies. The consultation was open to all interested stakeholders from both the public and private sectors, citizens, and representatives of the academic community.

This chapter presents different ways of data standardization and discusses the solutions established in some European countries. Finally, the standardization solution proposed by the Ministry of the Interior of the Czech Republic is described in detail in the last part of this chapter.

4.1 Ways of standardization

As previously stated, the path to greater interoperability leads through standardization. However, there are multiple ways to do that. This section aims to briefly introduce the three main standardization approaches based on the level at which they operate, particularly metadata, file format, and schema standards.

4.1.1 Metadata standards

The first step toward better open data quality is ensuring consistency of metadata [10], which is defined as *data that provides information about other data*[1]. In general, metadata may include various kinds of information, such as the author of the data, when and how the data was

1. `https://www.merriam-webster.com/dictionary/metadata`

collected, or the context for correct data interpretation and its effective re-use. Thus, different ways of describing datasets are problematic as they increase the chance of data misunderstanding or may even entirely hinder data discovery, and vice versa. With more standardized metadata, it is easier for both users and applications to find the correct data using only its metadata description [28].

The European Union ensures metadata consistency by utilizing European standards, such as the DCAT Application profile [10]. DCAT Application profile offers a uniform specification and guidelines for describing public sector datasets across Europe, allowing data portals to share dataset descriptions [29]. And DCAP-AP serves as the basis for the Czech standard DCAP-AP-CZ, which specifies the details of the data model for cataloging records in NKOD [30].

4.1.2 File format standards

Selecting the appropriate format, that is, how the data is structured and made accessible to the public, is critical for maintaining data usability and access [31]. However, different data formats are suitable for different kinds of data, and the context in which the data is used needs to be considered, too. Besides that, some data formats might not be widely available or widely understood, which may hinder their adoption and thus minimize the re-use potential. Therefore, in order to maximize data re-use, data providers may have to consider other file formats as well, depending on the users' demands, the context, and the nature of data.

The data can be represented in a number of different file formats, while each file format has its own set of constraints for data organization. This is a large amount of variation in and of itself. In order to reduce such variability, technical standards for the preferred data formats have emerged. Particularly, the technical standards issued by the Czech Ministry of the Interior [32] specify details regarding the structure, representation, or encoding of various kinds of data published in NKOD.

4.1.3 Schema standards

While the previously introduced standards operated rather from the outside of a dataset, schema standards help ensure consistency and data validity inside the dataset, too. These validations are typically conducted by comparing a dataset against a JSON Schema[2] [33].

A schema is a file describing the exact way the data should be formatted; it specifies the structure, content, and semantics of entities in datasets on a selected topic. Besides that, such a schema may also be used to ensure that additional validation constraints are met, like the phone number or an email address format.

Multiple European countries have started implementing their own schema standards in order to efficiently combine data, either from various data providers or across different areas; the British[3] or French[4] standards directories can be given as an example. In the Czech Republic, schema standards are in their early stages of development, covered by so-called *Open formal standards*. Because Czech schema standards represent the main focus of this work, we delve deeper into this topic in section 4.2.

However, it is important to note that standards cannot be declared on their own; a standard is determined based on its popularity. In case no one follows the developed schemas to publish their data, the data quality does not improve, and the developed tool is not useful. This most likely occurred in the case of Britain, where the most recent commit in the official schema GitHub repository[5] dates back three years ago, indicating the standards are no longer maintained nor developed. French open data community points out the same issue in their article [34]. Therefore, in addition to their development, it is also necessary to take measures to ensure that the standards are accepted by the public and used in practice.

2. `http://json-schema.org`
3. `https://schemas.opendata.esd.org.uk`
4. `https://schema.data.gouv.fr/schemas.html`
5. `https://github.com/esd-org-uk/schemas`

4.2 Open formal standards

Open formal standards (OFNs, in Czech: *Otevřené formální normy*) are technical guidelines created for selected domains. They aim to simplify data usage and ensure data interoperability, even when various data providers provide the same kind of data. Full interoperability is ensured in the following dimensions [13]:

- **Technical**: Covers the usage of standard web technologies for data identification.
- **Syntactic**: Deals with the specification and adherence to specific data formats, communication protocols and other system features that allow access to data from various sources in a standard way.
- **Semantic**: Specifies methods for describing and mapping data items, uniform vocabulary, and uniform classification of terms.

The assurance of interoperability represents the main reason why these standards should be adhered to if a publisher publishes data relevant to what OFNs are modeling. Apart from that, OFNs are binding on open data publishers according to Act No. 106/1999 Coll. on Free Access to Information.

4.2.1 Development process

OFNs are being developed and maintained through a collaborative decision-making process, which is coordinated by the Ministry of Interior of the Czech Republic. The development process takes place on the GitHub platform, so their creation may be tracked easily, and their final form can be even actively affected by data suppliers, municipalities, or even the general public. Currently, the GitHub repository[6] contains seven completed OFNs for particular datasets and two general OFNs, while OFNs for another seven domains are still under development [35].

6. `https://github.com/opendata-mvcr/otevrene-formalni-normy`

4.2.2 OFN features

Each standard describes how data about a specific topic, such as tourist points of interest, official board, or events, should be published. The description of OFN features is provided in the following paragraphs.

Content and form Each OFN begins with a description of essential terms for a given dataset which unifies the semantics of data; that is how the data is understood. The terms are represented in the form of a conceptual scheme, which models the terms as classes, their properties, and the relationships between them [36]. The scheme is also illustrated graphically for easier understanding. An example of such a standard in the form of a diagram is provided below in Figure 4.1. This format is uniform for all issued OFNs.

Shared specifications One of the key aspects of OFNs is that classes that appear in several OFNs, such as *Person*, *Contact*, or *Location*, are specified in a single place, so-called shared specifications. Currently, there are completed 11 specifications published in the GitHub repository [35]. In Figure 4.1, classes that occur in multiple OFNs are indicated by a blue frame.

Shared specifications ensure compatibility between the same entities in different datasets and thus facilitate data processing according to various OFNs. For example, information about admission to a concert provided in a dataset adhering to OFN for *Events* is represented in the same way as admission to a castle published in a dataset regarding *Tourist points of interest*. Furthermore, the use of shared specifications ensures at least partial compatibility between published datasets for which the OFN has not yet been created [36].

Generalization OFN also utilizes a common conceptual modeling tool known as generalization, represented in the scheme by a bond between two classes with a triangular arrow at one end.

Generalization example can be seen in the provided OFN scheme in Figure 4.1. General properties shared by multiple classes, such as identifier or name, are described in the class *Thing*. More specific classes, such as *Tourist point of interest* in our example, further specify the general class *Thing* and acquire its properties.

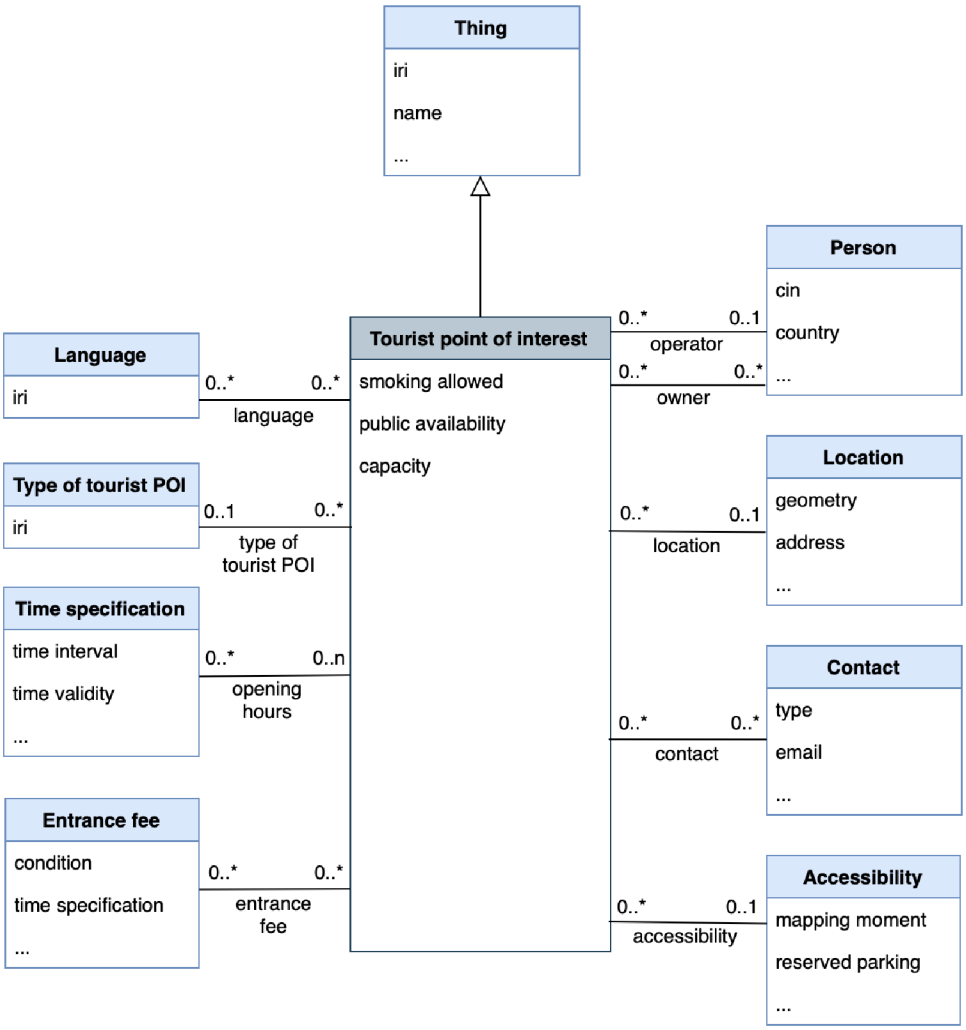

Figure 4.1: An example of OFN scheme for objects regarding *Tourist points of interest*. Translated to English [37].

```
{
    "title": "Turistický cíl",
        ...
    "properties": {
        "kouření_povoleno": {
            "type": "boolean",
            "title": "Kouření povoleno",
            "examples": [ true, false ]
        },
        "umístění": {
            "$ref": "https://ofn.gov.cz/umístění/2020-07-01/
            schémata/umístění.json"
        },
            ...
    }
}
```

Figure 4.2: A snippet of the JSON schema for the sample OFN *Tourist points of interest*.

Data format OFN ensures full syntactic interoperability by unifying the data format and specific data structures. In our example, OFN specifies that tourist points of interest should be provided in JSON-LD format [36]. Besides that, it determines a particular JSON structure for the data representation in the form of a JSON scheme, which can also be used for data validation.

Figure 4.2 represents a snippet from the JSON schema for our sample OFN, showing how the structure is defined for different properties. For example, the property *smoking allowed* (in Czech: *kouření povoleno*) has a specified data type, exact name, and a list of possible values. Property *location* (in Czech: *umístění*) refers to the *Location* class defined in shared specifications. The complete JSON schema is provided in Appendix A.

4.2.3 Current state of OFN adoption

Even though OFNs show a great deal of potential and are legally binding on data publishers, their actual application in practice is unsatisfactory. We checked the current compliance with OFNs regarding

Tourist points of interest manually with the usage of Open Data National Catalogue API[7], where we searched for the datasets having the *Tourist points of interest* standard's identifier given as a value of any attributes in the dataset's metadata, as defined by the OFN. The results were surprising. None of the published datasets regarding this topic has adhered to this standard yet, even one and a half years after its first official release[8].

There are various explanations for this phenomenon. First of all, data providers may be unaware that an OFN for their data type exists. At this time, datasets in NKOD are not automatically checked against these standards when being uploaded, so the data providers are not notified in case the dataset they are about to publish does not meet the OFN requirements. Another reason for their non-compliance may be the lack of motivation on the data provider's side to do so [38], for example, when they do not see the added value of publishing the data in a standardized form.

4.2.4 Motivation techniques

Publishing open data in a standardized format can improve the effectiveness, ease, and thus also the cost-effectiveness of future re-use of such data. However, as mentioned in the section above, one of the reasons for the insufficient adoption of OFNs is the lack of motivation on the data provider side. Without seeing the benefits and added value, modifying the data format is considered only unnecessary extra work. At the same time, it is often difficult for them to distinguish the open data itself from the application that works with this data [38].

To help solve these issues, a demo application[9] was developed. This application aims to demonstrate the impact of opening data straight into the real-world environment. It shows how a single general solution may be utilized for the same purpose in a variety of towns, in particular for creating leaflets with interactive design modifica-

7. https://data.gov.cz/fct/facet.vsp
8. The OFN *Tourist points of interest* was officially released in July 2020. `https://github.com/opendata-mvcr/otevrene-formalni-normy/tree/master/turistické-c%C3%ADle/2020-07-01`
9. `https://opendata-mvcr.github.io/app-ofn-plakaty/#/?iri=https%3A%2F%2Fwww.klaster-plasy.cz%2F`

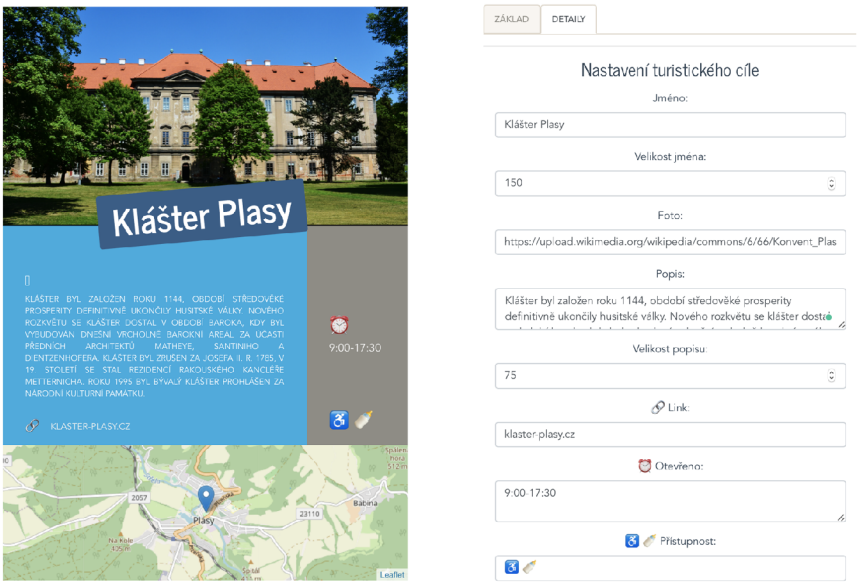

Figure 4.3: Example of a leaflet (left) generated by the demo application according to the given data (right)

tions, informing about tourist points of interest and sports venues, as displayed in Figure 4.3. The procedure of generating such a leaflet is as follows:

1. Municipalities provide data in a standardized form and register it in the Open Data National Catalogue, as required by law.
2. Because the data is published in the expected format, the application is able to process it automatically when prompted.
3. Based on the published data, municipalities can automatically create a leaflet with an overview of tourist points of interest or sports venues in their area.

Besides that, as discussed at the conference *Open Data 2021*, there is also an ongoing collaboration with *Mapy.cz*[10], a popular map portal in the Czech Republic. In the future, it is planned that this portal could

10. `https://en.mapy.cz`

automatically search NKOD for datasets adhering to a particular OFN and include the most up-to-date information about tourist points of interest in their maps.

Both examples above show how data from various providers can be automatically processed by different applications, very easily reused for other purposes, and create value. This, however, is conditioned by the correct usage of OFNs.

4.3 Summary

In the previous paragraphs, we have gradually introduced the concept of open data and its connection to smart cities. As the amount of published datasets has grown, the focus has switched; it has been confirmed that the quality of datasets plays a significant role in realizing open data's full potential.

In response to the need to increase the open data quality, the Czech Ministry of the Interior has developed a set of rules and standards, known as OFNs. However, the cell-level quality of published datasets is not verified in Open Data National Catalogue at this time and thus has been so far unknown. And therefore, multiple questions arise, in particular:

1. What is the actual cell-level quality of the datasets published in NKOD?
2. How to measure the quality of the datasets?
3. How are the standards adhered to? To what extent the existing datasets may be combined?
4. What are the most serious shortcomings of the datasets, and what are the recommendations toward improving their quality?

We will try to find answers to all these questions in the following chapters of this work.

5 Model proposition

The conference *Open Data 2021* revealed how many providers have their datasets' metadata problematic. However, since the idea behind open data is based on reusability and interconnection with other data, metadata is not the only aspect of data quality that should be evaluated. The current data quality checks in NKOD are not performed on the cell level, though.

The purpose of this chapter is to propose a framework that captures the aspects of data quality that are important to assess the current state of published datasets with respect to their interoperability. And since interoperability is ensured by the correct use of OFNs, the selected data quality dimensions are based on the rules set by OFNs. Because their adoption in the Czech Republic is still in its early stages, the five most fundamental data quality dimensions were chosen. Their extension by new dimensions is, of course, possible in future work.

5.1 Research approach

We took the intuitive approach [39] for this research; the selected data quality dimensions were chosen using an intuitive understanding of what characteristics are relevant for our case.

In terms of our domain focus, we chose datasets regarding *Tourist points of interest* for the data quality evaluation task. There are numerous reasons for the selection of such a topic:

1. The topic is relevant to the concept of smart cities development.

2. The OFN for the selected topic was officially released in July 2020 and belongs to the first set of published OFNs. Data publishers interested in this kind of data were therefore given a reasonable amount of time to study the standard and adapt their existing data structure and the actual content accordingly.

3. There is already an existing application for generating leaflets with tourist places and an ongoing collaboration with map portal *Mapy.cz*, as stated in subsection 4.2.4. Hence the benefits

of publishing data in a standardized form can be demonstrated in a very specific way.

5.2 Terms definition

Before we get into the specific dimensions, there are a few crucial concepts to understand. The following paragraphs aim to define terms used in the description of the proposed model.

Score Each dimension awards a certain number of points to a dataset based on its quality within the examined aspect. The given score ranges from 0 to 100 points, while the higher is the score, the higher is the dataset's quality. The calculated scores for individual dimensions remain separate and are not combined into one final score. Their purpose is to highlight the dataset's strengths and flaws in terms of interoperability and adherence to the particular OFN.

Feature As well as other OFNs for a specific dataset, OFN *Tourist points of interest* contains a minimal example specifying minimum requirements on modeled entities. If a dataset contains less information than provided in the minimal sample, the data is most likely meaningless, and no one will be able to use it [35]. Therefore, these minimal requirements are considered *mandatory* by the proposed evaluation model, even though all entities specified in the OFN schema are optional, according to the official documentation [40].

Besides the minimal example, OFN provides a more complex example, too. Such a complex model portrays elements that should be filled in to provide the user with the most accurate picture of the presented tourist destination. The additional information is considered *optional* by the proposed model.

Let denote the information contained at the dataset's top level as a *feature* of the *Tourist point of interest* object. The standard then specifies 7 mandatory and 13 optional features. Their complete list is provided in Table 5.1.

Table 5.1: List of mandatory and optional features and their weights for a dataset regarding *Tourist point of interest* specified by OFN

Type	Weight	Name
Mandatory	12.5	@context, typ, iri, název, umístění, vstupné, otevírací_doba
Optional	0.96	popis, kouření_povoleno, kapacita, veřejná_přístupnost, provozovatel, vlastník, kontakt, bezbariérovost, typ_turistického_cíle, jazyk, vytvořeno, aktualizováno, příloha

Feature weight Because each feature has different importance, their impact on the resulting score of the examined dimension should also vary. For example, missing information about the location of a tourist point can cause much more significant problems than missing input about languages spoken at the destination. Such importance of an individual feature is denoted as feature weight.

Feature weight is determined by its type. The weight of mandatory features is significantly higher since they carry the most critical information; their absence results in a considerable drop in the dataset's quality. In particular, the proposed model determines that the weight of one mandatory feature is equal to the sum of the weights of all optional features. Because the model operates on a scale ranging from 0 to 100, the weight calculation is performed as follows:

$$\Sigma\, w_{mf} + \Sigma\, w_{of} = 100 \tag{5.1}$$

$$w_{mf} = \frac{100}{N_{mf} + 1} \tag{5.2}$$

$$w_{of} = \frac{w_{mf}}{N_{of}} \tag{5.3}$$

where w denotes to a feature weight, N represents the number of features, mf stands for mandatory feature and of is obligatory feature. The calculated feature weights are displayed in Table 5.1.

Context In general, the proposed dimensions can be used to analyze the data quality in datasets from an arbitrary domain. However, the introduced model considers requirements defined by the standard for datasets regarding *Tourist points of interest*, to which the numbers in formulas are adapted. Therefore, coefficients such as feature weights may need to be modified in the future. This applies especially to situations when the evaluated data is outside our selected domain or when the evaluation model is used within a context, where the weight distribution of individual features differs from the suggested one.

5.3 Data quality dimensions

Our evaluation model consists of five fundamental dimensions for the data quality evaluation of open datasets regarding *Tourist points of interest*. Each of the proposed dimensions is introduced individually in separate sections of this chapter.

5.3.1 File format

Prior to combining diverse data sources, the data format is one of the key aspects that need to be considered [41]. Different data formats require different ways of data processing. Moreover, each data format places other requirements on the data structure, which in principle worsens data interoperability. There are three types of structures to consider, in particular tabular, hierarchical, and network [31].

- **Tabular**: Represents basic data structure, where the data is organized in rows and columns. The common data format for tabular structure is CSV.
- **Hierarchical**: Hierarchical data is a type of data format in which objects are connected to one another via parent-child connections in a tree structure. Typically, JSON data format or formats from the XML family are used for hierarchical data.
- **Network**: Relationships between any combination of items in any direction are possible with network structured data. An example of a file format enabling interlinking needed for network

creation is JSON-LD, which provides a way to increase JSON data interoperability throughout the web.

Data processing may be automated for many common formats. Nevertheless, the process gets more complicated if various data formats are used for the same data representation. It may even require manual assistance in case of inconsistencies caused by the nature of particular file formats, especially when also the data structure needs to be changed [42]. For example, it is necessary to deal with this issue if some data is published in a tabular format while others are organized hierarchically. Such a situation does not have a uniform and straightforward solution.

Formal requirements The OFN *Tourist points of interest* specifies that the data needs to be published in JSON format [37]. However, datasets related to tourist destinations published in NKOD use other formats, too, like CSV, XLSX, XML, or special formats from the JSON family, such as GEOJSON, causing issues for data to interoperate easily.

Score calculation The first dimension awards dataset a score based on its data format, and the score is given according to the conversion table displayed in Table 5.2. The individual levels of the proposed table have been designed based on the format's similarities to the required JSON format. The levels can also be understood to represent the ease of transformation from the particular file format to JSON.

Table 5.2: Conversion table for awarding scores based on file format

Score	File format
100	JSON, JSON-LD, GEOJSON
75	XML, GML, KML, RDF
50	CSV
25	XLS, XLSX
0	PDF, TXT

As displayed in Figure 4.1, the object *Tourist point* itself is in relation with multiple objects such as *Person*, *Location*, or *Contact*, while each of them has its attributes and can even include other nested objects. This indicates that a format allowing hierarchical structure is needed. Therefore, tabular file formats allowing only flat design, such as CSV, are penalized to some extent. A detailed explanation for the individual groupings is provided below.

- **100 points**: The highest score is given to file formats from the JSON family, such as JSON, JSON-LD[1], or GEOJSON[2]. Even though a valid GEOJSON data structure differs from the data structure specified by OFN for *Tourist points of interest*, syntactically are GEOJSON files valid JSON files and therefore fully meet the requirements defined by the OFN in terms of file format. Requirements in which GEOJSON files may cause interoperability issues, such as names of individual attributes, are checked in other proposed quality dimensions, focusing on data representation. Similar to GEOJSON, JSON-LD file format is a valid JSON format, too. It just facilitates data interoperation through special keys like *@context* [43]. Example datasets[3] regarding *Tourist points of interest* showing the correct usage of OFNs adopt the *@context* key for interlinking the dataset with the corresponding standard.

- **75 points**: The second-best score is awarded to datasets with file format from the XML family[4]. XML documents allow hierarchy in the same way as JSON. Besides that, the data format can be generally changed to JSON quite easily. There are numerous online tools for automatic conversion of XML files to JSON structure, such as *ConvertJson*[5] application. However, a potential risk may be posed by semi-structured elements, which are

1. JSON-LD file format is used for encoding linked data: `https://json-ld.org`
2. GEOJSON is a file format used for encoding geographic data structures: `https://geojson.org`
3. Example of a valid dataset according to OFN *Tourist points of interest*: `https://github.com/opendata-mvcr/otevrene-formalni-normy/blob/master/turistické-c%C3%ADle/2020-07-01/př%C3%ADklady/2.jsonld`
4. XML documentation: `https://www.xml.com`
5. `https://www.convertjson.com/xml-to-json.htm`

```
<e>
    a semi-structured
    <a> element </a>
</e>
```

Figure 5.1: Example of a semi-structured XML element

composed of a mix of text and child elements [44]. These elements may cause issues in preserving data semantics when converting an XML file to JSON format. An example of such XML data is provided in Figure 5.1.

- **50 points**: Datasets with a tabular structure, such as CSV files, are awarded 50 points as the file format does not allow hierarchical structure and therefore is not suitable for complex data. However, the hierarchy can be possibly realized by publishing multiple CSV files, where each file represents a separate modeled object. Another way of maintaining some kind of hierarchy is by publishing a single file with a denormalized data structure. In any case, both solutions add extra work to the data processing phase and obstruct data in smooth interoperation.
- **25 points**: The second-lowest score is given to datasets with tabular structure that are published in a proprietary[6] file format like XLS or XLSX. These formats are problematic because special software may be required to read data properly, which contradicts the general concept of open data. The need for use of a proprietary software unnecessarily increases the costs and adds extra work to the data processing phase, especially when converting XLS files to the non-proprietary CSV format is a trivial operation.
- **0 points**: No points are given to files published in a format that does not guarantee any structure, for example, TXT or PDF. Files like these are machine-unreadable, and in such a case, manual processing is needed for every single file.

6. https://opendatahandbook.org/glossary/en/terms/proprietary/

5.3.2 Schema accuracy

Schema accuracy refers to the syntactic accuracy of features' names compared with the naming convention defined by OFN. The focus on the naming of modeled entities is an essential part of quality checks with an emphasis on interoperabilty, as two entities cannot be merged automatically if their names are different, even if they semantically represent the same object. A real example of incorrectly named features can be, for instance, using English feature names such as *location* instead of the Czech word *umístění*.

Formal requirements The standard regarding *Tourist points of interest* specifies 7 mandatory and 13 optional features, as displayed in Table 5.1. Except for the unique key *@context*, specified by JSON-LD format [43] and used for interlinking the dataset with the corresponding standard against which a dataset structure is valid, all feature names are in the Czech language. There is always only one correct feature name.

Score calculation The dataset's score in terms of schema accuracy is calculated based on the correctness of individual feature names. The correctness of the naming of mandatory and optional features impacts the score in various ways, as presented in the Equation 5.4 below:

$$score = \sum_{f=1}^{F} w_f n_f \begin{cases} n = 1, & \text{if } n \in N \\ n = 0, & \text{otherwise} \end{cases} \tag{5.4}$$

where f denotes a feature specified by the standard, w_f is the weight of a feature, and n_f represents the name of a feature. A set of both compulsory and optional features specified by OFN are denoted by F, while N represents the set of all feature names contained in the examined dataset.

The minimal example containing all mandatory features that are correctly named achieves a score of 87.5 points; the full score of 100 points is given to datasets containing precisely named both compulsory and optional features. Therefore, all datasets scoring above the threshold of 87.5 points can be considered of high quality.

5.3.3 Schema completeness

Schema completeness is dedicated to checking the semantic correctness of the information carried by individual features. In other words, the dimension focuses on the features' content, regardless of the correctness of their naming. Data types used are also ignored in this case. In particular, if all information defined by the standard is contained in the dataset, the resulting score is high and vice versa; the score decreases in case some features are missing, especially those which are mandatory.

Before the actual score calculation, each examined dataset needed to be manually adjusted. This means that the original features' names had to be changed to match the words specified by OFN, but only if the information carried by the feature semantically matched the standard. Subsequently, these altered features' names were checked against the standard to get the score. An example is provided below in Figure 5.2 and Figure 5.3.

Formal requirements The standard defines that a dataset concerning *Tourist points of interest* must contain data about their type, iri, name, location, opening hours, and entrance fee. The information referring to the used standard is also required.

Besides that, the extension of the dataset by providing additional information is possible. The standard specifies that supplementary data may include information about the object's owner, capacity, or contact information. The complete list of optional information is displayed in Table 5.1.

Score calculation The score for schema completeness is determined in a similar way as for the schema accuracy dimension. However, the dimension's focus is different. This time, the score is based on the presence or absence of expected information contained in a dataset. The formula is given below.

$$score = \sum_{f=1}^{F} w_f i_f \begin{cases} i = 1, & \text{if } i \in I \\ i = 0, & \text{otherwise} \end{cases} \tag{5.5}$$

where f denotes a feature specified by the standard, w_f is the weight of a feature, and i_f represents the information carried by a feature. A set of both compulsory and optional features specified by OFN is denoted by F, while I represents the set of the expected information contained in the dataset according to the standard. The weights remain different for mandatory and optional features. They are provided in Table 5.1.


```
{
    "ObjectId": 26,
    "name": "Brněnské výstaviště",
    "text": "Areál brněnského výstaviště patřil ... ",
    "url": "https://www.gotobrno.cz/misto/brnenske-vystaviste/",
    "address": "Výstaviště 405/1, 603 00 Brno, Česko"
}
```


Figure 5.2: Sample data about a tourist destination in its original form. The information about the object's *name*, *description*, *location* as well as the *identifier* are contained in the dataset but their naming is incorrect. This causes that the presence or absence of such features cannot be inspected automatically.


```
{
    "ObjectId": 26,
    "název": "Brněnské výstaviště",
    "popis": "Areál brněnského výstaviště patřil ... ",
    "iri": "https://www.gotobrno.cz/misto/brnenske-vystaviste/",
    "umístění": "Výstaviště 405/1, 603 00 Brno, Česko"
}
```


Figure 5.3: Features' names which semantically matched the information specified by the standard were renamed. Features *název*, *popis*, *umístění*, and *iri* now correspond with the standard, and therefore, the presence of expected features can be checked against the OFN. However, the dataset lacks data about the opening hours or entrance fee, which lessens the dataset's score in terms of schema completeness.

5.3.4 Data type consistency

The quality of inputs determines the ability to fuse digital data and create relevant information [45]. Because the same kind of data can be produced in various formats and using different data types for the same data requires time-consuming data cleansing, it is suitable to monitor the aspect of data consistency. In particular, this dimension focuses on the consistency of data types used for the same data representation within a feature.

Data of low quality in terms of data consistency may not immediately imply a major concern. If data within an optional feature is not represented consistently, the information can be omitted without causing substantial harm. However, in the case of data inconsistency within mandatory features, the consequences are more severe, potentially resulting in completely restrained data usability. This is why different weights for compulsory and optional features need to be applied for score calculation.

The implementation of the proposed model can distinguish five primary data types, namely *integer*, *float*, *bool*, *string*, and *null*. On top of that, we extended the type recognition by five custom data types, which were selected based on the analysis of the actual values provided in the examined datasets. The list of the custom data types is as follows:

- **URL**: recognized by a function for URL recognition contained in a python library *validators*,
- **E-mail**: recognized by a function for e-mail recognition contained in a python library *validators*,
- **Address**: identified by a regex string searching for an address written in Czech format consisting of *street name* and *number* in given order,
- **Point**: recognized by a pattern, where geographical coordinates are wrapped by the *POINT* label,
- **Phone number**: identified by a regex string searching for a phone number, written in Czech or international form.

Formal requirements All values within one feature should be represented by the same data type.

Score calculation The score is affected by the number of used data types within one feature and its weight. This time, the dimension focuses on the values themselves rather than on their comparison against the standard's schema. Because of that, the earlier-defined explanation of mandatory and optional features needs to be slightly adjusted. The definition of individual feature types for this dimension is as follows:

- **Mandatory features**: Features defined as mandatory by the standard, which are at the same time contained in the examined dataset.
- **Optional features**: All other features which are included in the examined dataset.

In other words, the feature weights are derived from the total number of provided features within a dataset, not solely from the features defined by OFN. Suppose a dataset containing two mandatory features out of seven defined by the standard, and zero optional features. Then the weight of each feature is 50. Compulsory features which are missing have no effect on the score calculated for this dimension.

This means that even if a dataset is missing some mandatory features and does not include any optional ones, it can still score a maximum of 100 points in terms of data consistency. Naturally, all values within each feature must be represented by the same data type in such a case, otherwise the score gets lower.

Multiple data types used within a single feature are stringently punished. The actual formula for score calculation is provided below.

$$score = \sum_{f=1}^{F} w_f \frac{1}{t_f} \tag{5.6}$$

In Equation 5.6, f denotes a feature provided in the dataset, w_f is the feature weight, and t_f represents the number of data types used within a feature. A set of all features provided in the dataset is denoted by F.

5.3.5 Data completeness

The amount of data we collect is growing. However, not all of the data is complete, and some information can be missing. Missing values are typically denoted with a null value, a specific mark indicating that a value is absent or undefined [46], but empty strings are also widely used.

As stated in the literature, there are two main reasons for providing incomplete data [46]. The information may be unknown to the data source, or it refers to a property that is not relevant to the particular object. In any case, null values represent a severe quality issue. They bring confusion to further data processing, may lead to wrong data interpretations, or completely restrict data to be usable. Therefore, the data processor needs to be aware of them in order to choose the right strategy for their correct handling.

Formal requirements All values within a feature should ideally be complete; that is different from null and empty objects.

Score calculation The data completeness dimension measures the weighted ratio of non-null values. By ratio, we mean the division of non-null values by all values provided within the examined feature. Both values of data type null and empty strings are understood as null values in the proposed model.

As for the weights, their calculation process is the same as described in subsection 5.3.4 for the data consistency dimension; the total number of given features within a dataset, and not only the features described by OFN, is used to calculate feature weights. The formula is as follows:

$$score = \sum_{f=1}^{F} w_f \frac{n_f}{v_f} \tag{5.7}$$

where w_f is the weight of a feature, n_f represents the number of non-null values within a feature, and v_f denotes the number of all values within the examined feature. A set of both mandatory and optional features provided in the dataset is marked as F.

6 Evaluation

We used the evaluation model introduced in chapter 5 to measure the data quality of Czech open datasets on *Tourist points of interest*, with focus on their interoperability and their adherence to the standards developed by the Czech Ministry of the Interior. Because the data quality of datasets published in NKOD is currently monitored only on the metadata level, their cell-level quality was so far unknown.

This chapter presents the results of the datasets' evaluation in terms of five fundamental dimensions, namely *File format*, *Schema accuracy*, *Schema completeness*, *Data type consistency*, and *Data completeness*. Besides that, the chapter also provides a detailed analysis of the results separately for each dimension and presents the reviewed datasets in a more detailed manner.

6.1 Datasets

The publication of data on tourist attractions is not as widespread as initially expected; nowadays, only a few data sources provide such information. Despite our extensive search even outside NKOD, we did not manage to obtain more data regarding our selected topic. This finding may also imply that municipalities must continue to be supported in their efforts to publish open data in order to take full advantage of their benefits.

However, datasets outside NKOD were usually unusable for our case. They were of a slightly different topic than tourist attractions, for example, a dataset from the Nový Jičín municipality regarding cultural organizations[1]. Other issues concerned the data availability, such as in the open data portal of Plzeň municipality. Even though the dataset[2] seemed to provide the required kind of information, the data was integrated into an application but not available for download as raw data.

Overall, we managed to collect 14 datasets from 6 different municipalities, all published in NKOD. Because the quality of datasets from

1. `https://www.novyjicin.cz/opendata/`
2. `https://opendata.plzen.eu/public/opendata/detail/97`

the same provider was essentially the same, we decided to analyze only one dataset from each provider so that the results were unbiased. Therefore, the evaluation concerns six datasets from six distinct municipalities, which are listed in Table 6.1.

Each of the selected datasets provides a different number of records and features. The overview is listed in Table 6.2. While the dataset from Huntířov municipality contains 10 records, Brno offers a slightly more rich collection of 350 tourist destinations. As for the features, both Praha and Brno datasets capture 9 pieces of information regarding each record. The largest number of 71 features offers datasets from Děčín. However, it is necessary to note that many features in this dataset do not contain any meaningful value other than null.

Table 6.1: List of municipalities providing data regarding *Tourist points of interest* and selected datasets for analysis

Municipality	Number of relevant datasets	Selected dataset
Brno	2	Turistická místa [a]
Děčín	1	Seznam bodů zájmů (POI) [b]
Hradec Králové	8	Zámky [c]
Huntířov	1	Turistické cíle [d]
Ostrava	1	Turistické cíle [e]
Praha	1	Významné vyhlídkové body [f]

a. `https://data.gov.cz/zdroj/datové-sady/44992785/feeac4ed9a63085ff023c5057d5ddf0d`
b. `https://data.gov.cz/zdroj/datové-sady/https---opendata.mmdecin.cz-api-3-action-package_show-id-seznam-bodu-zajmu-poi`
c. `https://data.gov.cz/zdroj/datové-sady/70889546/dcf391b507e3810732690e3414bbe71c`
d. `https://data.gov.cz/zdroj/datové-sady/00261360/3746578`
e. `https://data.gov.cz/zdroj/datové-sady/https---opendata.ostrava.cz-api-3-action-package_show-id-turisticke-cile`
f. `https://data.gov.cz/zdroj/datové-sady/http---opendata.praha.eu-api-3-action-package_show-id-ipr-vyznamne_vyhlidkove_body`

Table 6.2: Overview of the number of records and features provided by the selected datasets

Municipality	Name of the selected dataset	Number of records	Number of features
Brno	Turistická místa	350	15
Děčín	Seznam bodů zájmů (POI)	254	71
Hradec Králové	Zámky	33	9
Huntířov	Turistické cíle	10	18
Ostrava	Turistické cíle	66	18
Praha	Významné vyhlídkové body	323	9

6.2 Evaluation results

We analyzed each dataset in terms of the five proposed data quality aspects. In each dimension, a dataset could get a score ranging from 0 up to the maximum of 100 points. The full record of achieved scores in each data quality dimension are provided in Table 6.3. A better overview of the results for individual municipalities is then visualized by radar charts in Figure 6.1.

At first examination, it is evident that all datasets have significant deficiencies in the dimension of schema correctness. This observation is also confirmed by Figure 6.2, where all the scores for each data quality dimension achieved by individual datasets are shown on top of each other. In section 6.3, all results findings are analyzed in further depth.

Table 6.3: Complete results of earned scores in five monitored dimensions for individual municipalities

Municipality	File format	Schema accuracy	Schema compl.	Data type consistency	Data compl.
Brno	100	0	65.38	53.49	64.46
Děčín	25	0	53.85	58.03	75.99
Hr. Králové	100	0.96	38.46	97.94	100
Huntířov	50	0	52.88	70.97	76.3
Ostrava	50	0	52.88	65.73	81.03
Praha	100	0	25	78.56	100

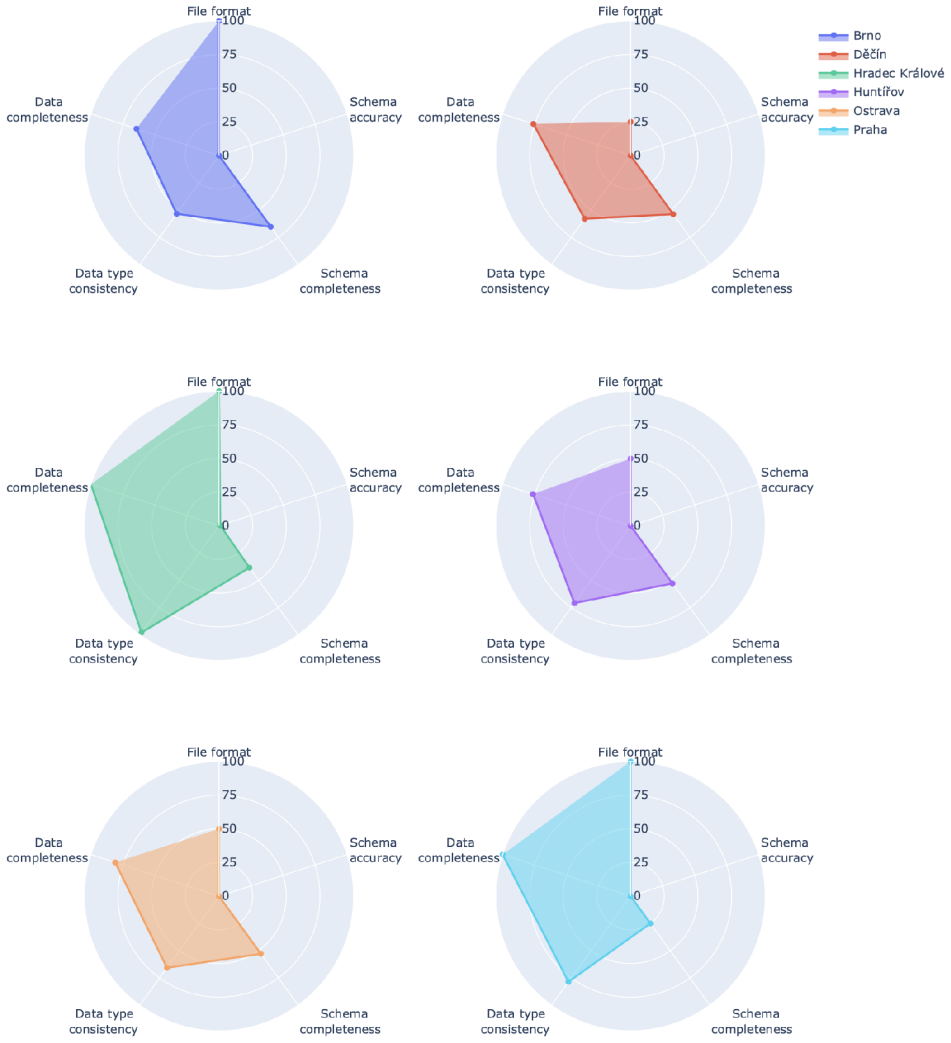

Figure 6.1: Radar charts visualizing the achieved scores in each data quality dimension for individual municipalities

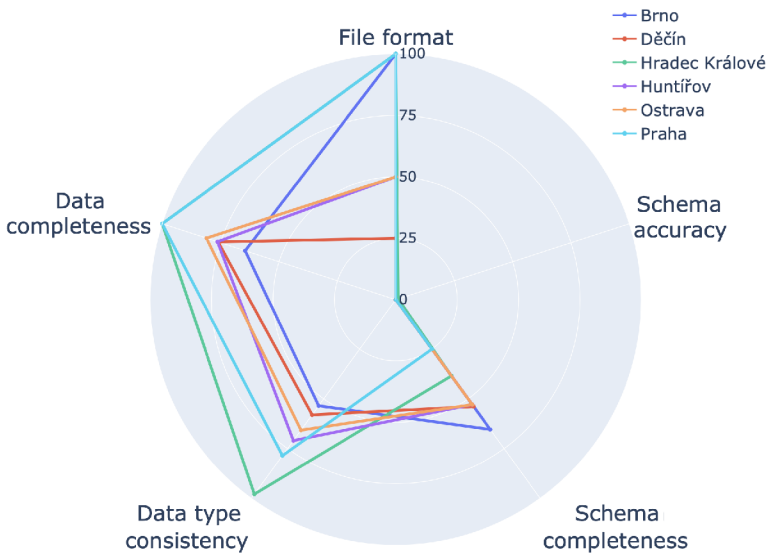

Figure 6.2: Comparison of the achieved scores in each data quality dimension for individual municipalities

6.3 Results analysis

A statistical overview of the datasets' individual scores aggregated by dimension is provided in the box plot displayed in Figure 6.3. The findings are diverse; while datasets generally achieve a somewhat satisfactory score in some dimensions, such as *Schema completeness*, other aspects, especially *Schema accuracy*, prove significant deficiencies for all datasets. The following section analyses the retrieved results separately for each dimension in a detailed manner.

6.3.1 File format

The results of the file format dimension turned out to be the most diverse one in terms of the achieved scores. Three datasets (Brno, Hradec Králové, Praha) scored the maximum number of points for using a file format required by the standard, and two datasets (Ostrava, Huntířov) earned 50 points as they provided the data in CSV format. The score of Děčín municipality is only 25 points as their data was

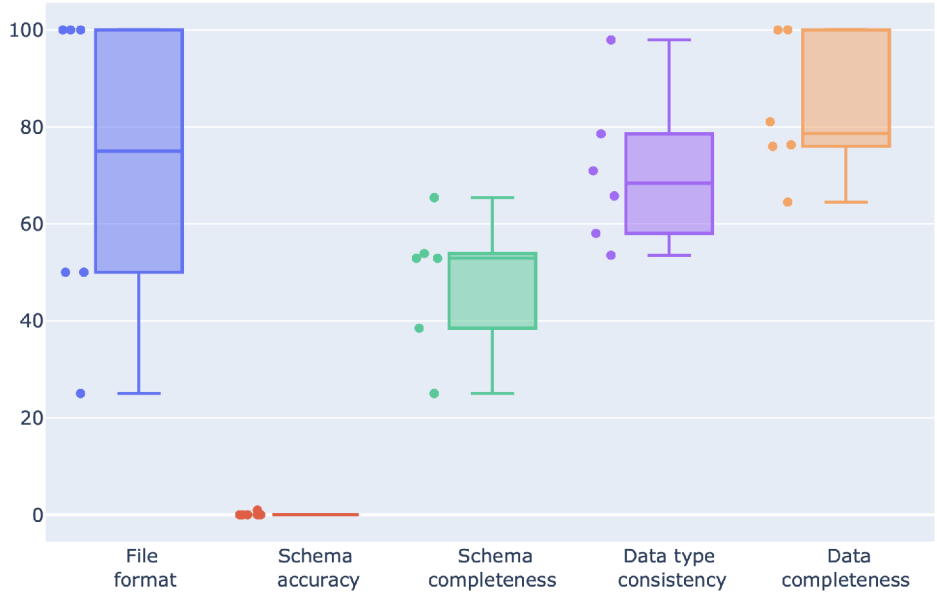

Figure 6.3: Statistical overview of the achieved scores aggregated by the proposed data quality dimensions

initially published as an XLS file, which is a proprietary file format and thus contradicts the general concept of open data.

If we focus solely on the data structure of the examined datasets, half of the data was published in hierarchical, while the other half was in tabular form. And this is a fundamental issue that needs to be addressed in the first place as it represents a major obstacle in data interoperability. Not only do different data structures require different ways of data processing, but they also use different ways of data representation, which may lead to further issues.

The structure of CSV files is easier, and thus, the format is more comfortable to work with, which explains their popularity among ordinary people. However, they do not directly support data hierarchies, which applies the used OFN to process more complex data, for which tabular structure is inconvenient. As a result, combining datasets with different structures is challenging, as non-trivial data processing is needed. Format unification across datasets according to the standard is indeed desirable for data to fully interoperate.

6.3.2 Schema accuracy

The results from the dimension monitoring schema accuracy have turned out to be the poorest of all the dimensions studied. Five out of six examined datasets scored 0 points as none of the feature names corresponded with the standard. One dataset was awarded 0.96 points for using a correct name in case of one optional feature.

There are multiple reasons why all datasets achieve unsatisfactory results. After a thorough review of the feature names, we identified the following causes:

- two datasets (Brno, Hradec Králové) use English feature names,
- two datasets (Huntířov, Ostrava) use Czech feature names, but without diacritics or modified in any other way,
- two datasets (Děčín, Praha) use a combination of both English and Czech feature names, even without diacritics or modified in any other way.

We observed that each municipality uses its own set of unique feature names. However, certain similarities can be seen in the datasets. For example, the datasets from Ostrava and Huntířov share 10 out of 18 feature names, even though they differ from feature names required by the standard. Such a match is certainly not accidental. Besides that, it is suspicious, especially when these two municipalities are located on opposite sides of the Czech Republic. Therefore, we believe such an observation indicates there must have been some kind of collaboration between those two municipalities, resulting in a dataset structure that is roughly half the same. In any case, this is the path on which the situation should proceed because constant data exchange will only emphasize the need for the same naming conventions applied in published datasets.

To sum up, consistently applied naming practices are crucial to ensuring data interoperability. As two entities representing the same object but with different names cannot be effortlessly combined, joining any two of those datasets would be impossible in their raw form. Thus, the feature naming in the examined datasets should be changed according to the rules specified by the standard. The interoperability can be ensured only if the naming is united across datasets.

6.3.3 Schema completeness

Schema completeness compared the presence of information carried by features contained in the dataset against the standard, regardless of the correctness of their naming and data type. After manual changing of the features' names, we found out that even despite inaccurate feature naming, the examined datasets do contain the information required by the standard, at least to some extent.

The highest score of 65.38 points was achieved by the dataset published by Brno municipality. On the other hand, the Prague dataset scored the lowest score of 25 points, which corresponds with merely two mandatory features being included. The median score is equal to 52.88 points.

As for the most commonly present and absent information carried by mandatory features, only information about the *name* (název) and the *location* (umístění) of the listed tourist points of interest appeared in each analyzed dataset. The representation of *iri* identifiers is likewise reasonably sufficient. On the contrary, none of the datasets contained information about the tourist attraction's *type* or the dataset's *context*. However, a complete absence of *@context* information is expected because this feature is intended for providing a link to the used standard. And as stated in subsection 4.2.3, the adoption of OFN for datasets regarding *Tourist points of interest* is currently unsatisfactory. A detailed overview of the results in terms of schema completeness is visualized in a bar chart in Figure 6.4.

The fact that the datasets are often provided as a list of places focused simply on their location was recognized as the primary source of the observed difficulties in terms of schema completeness. This is especially true for datasets published in GEOJSON format, designed specifically for this purpose. But even a dataset that might be of sufficient quality for one purpose may not be suitable for another [17]. Therefore, it would be desirable to append the missing information in order to increase the possibilities of employing this data in the tourism industry. Information on opening hours and admission is required in the context of tourism, and the fact is that such information is currently lacking in at least half of the obtained datasets.

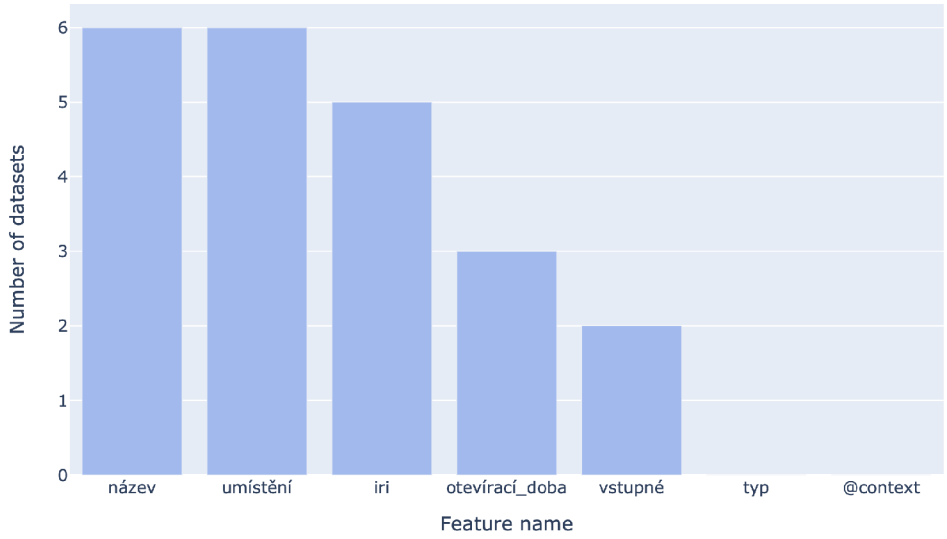

Figure 6.4: Overview of the most commonly present and absent information carried by mandatory features in the examined datasets

6.3.4 Data type consistency

According to the results presented in Table 6.3 and Figure 6.3, the median of the achieved scores reaches a value of 68.35 points in terms of data type consistency. Unambiguously the highest score of roughly 98 points was acquired by the dataset from Hradec Králové municipality. Values in this dataset are consistent within provided features to a large extent. The analysis of the overall results is given below.

The bar chart provided in Figure 6.5 shows the ratio of mandatory and optional features using various numbers of data types. In general, values within mandatory features turned out to be less consistent in their data type than optional ones. About one-third of mandatory features contain values of three or even four different data types within one dataset. Having so many distinct ways of representing data within one dataset is problematic for data interoperability. On the other hand, the results also show that approximately 45% of compulsory features and around 50% of optional features are consistent; they include values of the same data type.

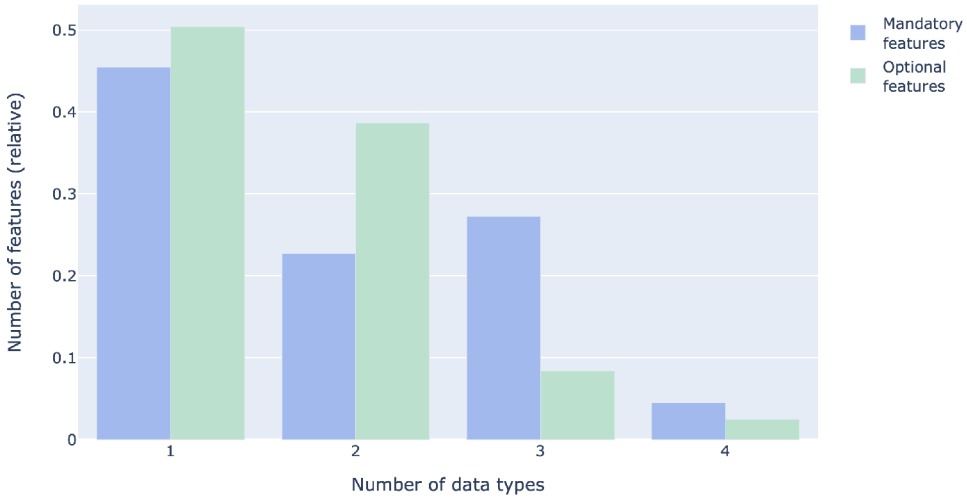

Figure 6.5: Bar chart visualizing the number of used data types by mandatory and optional features

6.3.5 Data completeness

Data completeness dimension measured the ratio of non-null values within the provided features. Based on the results presented in Table 6.3, it can be stated that all datasets scored relatively high. The median value of 78.66 points is the highest achieved value among all monitored aspects. What is more, two datasets (Hradec Králové, Praha) have been awarded the maximum of 100 points as they do not contain any null values or empty objects.

As for the detailed analysis of the ratio of non-null values within mandatory and optional features, the results are provided by a bar chart in Figure 6.6. Roughly 60% of all mandatory and 35% of optional features include only non-null values. Some null values or empty objects were discovered in another 30% of obligatory features; however, at least half of the values contained inside such features are meaningful. A significant predominance of null values was found within 9% of mandatory features. The completeness of optional features was found to be more problematic; more than 35% of optional features provided in the examined datasets contain a vast majority of null values or empty objects.

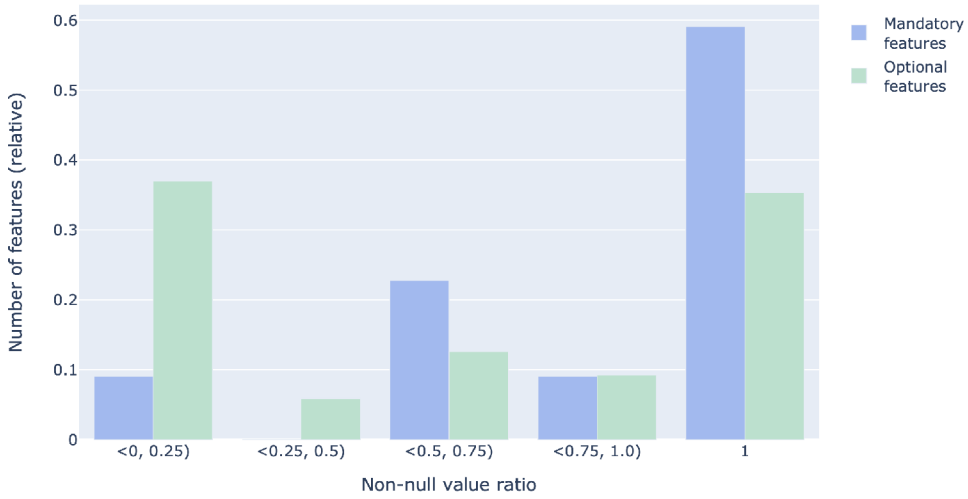

Figure 6.6: Bar chart visualizing the ratio of non-null values within mandatory and optional features

6.4 Summary

We proposed a fundamental evaluation model consisting of five data quality dimensions in the practical part of this work. The individual dimensions were chosen so that interoperability of datasets and adherence to specified standards were considered in the evaluation.

The model was used to assess the datasets' quality on tourist points of interest. The datasets were downloaded from the Open Data National Catalogue, and the evaluation's findings reveal their previously unknown cell-level quality. In particular, these are the most important findings of the assessment:

- The used file formats are diverse. Half of the datasets display the data in tabular form, which causes issues when modeling more complex data like this.
- Schema accuracy results turn out to be the poorest of all the dimensions studied; all datasets scored 1 or less points out of 100. This indicates that the datasets do not adhere to the standard in terms of feature naming conventions. However, certain similarities were found between some datasets.

- Regardless of the incorrect naming of features and their data types, around 53 % of the information required by the OFN was present in the datasets. Most dasasets model individual objects only as locales rather than in the context of tourism.
- The data types are often not united within the same feature and dataset; only 45 % of mandatory and 50 % of optional features contain values of uniform data type.
- Mandatory features are relatively complete. Around 70% of all compulsory features contained in the examined datasets include none or a few null values.

To sum up, the most severe shortcomings concern the datasets' compliance with the standard, causing issues in their interoperability. When data providers start publishing datasets according to the requirements specified by the corresponding OFN, their quality, as well as their potential to be re-used in tourist industry, will significantly increase. Otherwise, datasets with different feature names cannot be effortlessly combined. Particular recommendations for improving the datasets' quality are provided below, separately for each dimension in Appendix B.

7 Conclusion

The goal of this thesis was to examine the current state of open data standardization within Europe, with a particular emphasis on the measures implemented in the Czech Republic. Besides that, the thesis also aimed to provide a fundamental view on the so-far unknown cell-level data quality of Czech datasets on a selected topic published in the Open Data National Catalogue. As for the chosen topic, we decided to focus on the quality of datasets regarding Tourist points of interest.

Open formal standards, one of the standardization solutions developed by the Czech Ministry of the Interior, represented the core of this work. These standards, also known as OFNs, aim to increase open data quality and ensure better interoperability between datasets published by various data providers. As they are legally binding on data providers, we inspected their adoption and application in practice.

Despite an extensive search, we did not manage to find a single dataset adhering to the standard on the selected topic, based on the information provided in the metadata records. Therefore, we took only non-standardized datasets from six different data providers, which seemed to include the relevant information, and performed the cell-level quality assessment. Because such an evaluation has not been performed yet, we needed to design a way to measure the datasets' quality at first.

Our proposed model consists of five dimensions, namely *File format*, *Schema accuracy*, *Schema completeness*, *Data type consistency*, and *Data completeness*. Each dimension captures a different aspect of data quality that is essential for ensuring interoperability, which is also conditioned by the proper use of OFNs. As for the way of evaluation, each dimension awards a score ranging from 0 to a maximum of 100 points to a dataset.

The results are diverse. The examined datasets showed moderate deficiencies in terms of data completeness and data type consistency; that is in dimensions monitoring the ratio of non-null values and the number of data types used for the representation of the same kind of data within individual features in one dataset.

However, more important is that the results confirmed that the adoption of OFNs is unsatisfactory in the Czech Republic. The most

severe shortcomings were identified in the datasets' adherence to the standard, especially during the inspection of using proper naming conventions and including all required information. All datasets achieved less than 1 point in this dimension. Besides that, most datasets model individual items merely as pure localities and not as tourist objects as required. As a result, published datasets are practically incapable of interoperability in their current state in the tourism context, and their potential to create any value is minimal.

We identified various explanations for this phenomenon. Firstly, data providers may be uninformed of the existence of an OFN for their data type, as no such checks are performed during the process of publishing datasets in NKOD. Secondly, data providers may lack the motivation or skills needed to adjust their datasets to match the standard. It is possible they do not realize the importance of providing the data in a standardized format as they do not perceive any potential added value. And finally, another issue is the insufficient amount of published data in general. These are the challenges that need to be overcome in the future as a matter of priority.

7.1 Future work

There are several possible additions to this work. As our proposed model consists only of five fundamental dimensions, it may be appropriate to deal with its extension to new dimensions. The proposed model might also be used to assess the quality of data of another domain, ideally on a more considerable amount of data.

However, as stated above, the publication of open data is not sufficient, so creating a larger data bank may be impossible at this time. This leads us to the final option of looking into ways how to encourage data producers not only to publish their data in standardized form, but also to share more of their data in general. Only then can society make full use of the open data potential.

Bibliography

1. EU, European Data Portal. *The benefits and value of open data*. Publications Office of the European Union, 2020-01-22. Available also from: https://data.europa.eu/en/highlights/benefits-and-value-open-data.
2. ŠTĚPÁNEK, P.; GE, M.; WALLETZKÝ, L. IT-Enabled Digital Service Design Principles - Lessons Learned from Digital Cities. In: THEMISTOCLEOUS, M.; MORABITO, V. (eds.). *Information Systems*. Cham: Springer International Publishing, 2017, pp. 186–196. ISBN 978-3-319-65930-5. Available also from: https://www.springerprofessional.de/en/it-enabled-digital-service-design-principles-lessons-learned-fro/14215422.
3. HUOVILA, A.; AIRAKSINEN, M.; PINTO-SEPPÄ, I.; PIIRA, K.; PENTTINEN, T. Smart City Performance Measurement System. In: 2016. Available also from: https://www.researchgate.net/publication/313369674_SMART_CITY_PERFORMANCE_MEASUREMENT_SYSTEM.
4. BATAGAN, L. Open Data for Smart Cities. *Economy Informatics*. 2012, vol. 12, no. 1/2012, pp. 136–142. Available also from: http://www.economyinformatics.ase.ro/content/EN12/13%20-%20Batagan.pdf.
5. AHLGREN, B.; HIDELL, M.; NGAI, E. C. H. Internet of Things for Smart Cities: Interoperability and Open Data. *IEEE Internet Computing*. 2016, vol. 20, no. 6, pp. 52–56. Available from DOI: 10.1109/MIC.2016.124.
6. *Open Data in 60 Seconds*. The World Bank. Available also from: http://opendatatoolkit.worldbank.org/en/open-data-in-60-seconds.html.
7. *Open Definition 2.1*. Open Knowledge Foundation. Available also from: https://opendefinition.org/od/2.1/en/.

8. CARRARA, W.; ENZERINK, E.; OUDKERK, F.; RADU, C. et al. *Open Data Goldbook for Data Managers and Data Holders: Practical guidebook for organisations wanting to publish Open Data*. Publications Office of the European Union, 2018. Available also from: https://data.europa.eu/sites/default/files/european_data_portal_-_open_data_goldbook.pdf.

9. EU, European Data Portal. *What is open data*. Publications Office of the European Union. Available also from: https://data.europa.eu/en/trening/what-open-data.

10. BERENDS, J.; CARRARA, W.; ENGBERS, W.; VOLLERS, H. *Reusing Open Data: A study on companies transforming Open Data into economic & societal value*. Luxembourg: Publications Office of the European Union, 2020. ISBN 978-92-78-41872-4. Available also from: https://data.europa.eu/sites/default/files/reusing_open_data.pdf.

11. HUYER, E.; KNIPPENBERG, L. *The Economic Impact of Open Data: Opportunities for value creation in Europe*. Luxembourg: Publications Office of the European Union, 2020. ISBN 978-92-78-42108-3. Available also from: https://data.europa.eu/sites/default/files/the-economic-impact-of-open-data.pdf.

12. VAN HESTEREN, D.; WEYZEN, R.; LINCKLAEN ARRIENS, A.; BLANK, M. *Open Data Maturity: Report 2020*. Luxembourg: Publications Office of the European Union, 2020. ISBN 978-92-78-42406-0. Available also from: https://data.europa.eu/sites/default/files/edp_landscaping_insight_report_n6_2020.pdf.

13. *Veřejný datový fond*. Ministerstvo vnitra České republiky, 2021-09-06. Available also from: https://archi.gov.cz/nap:verejny_datovy_fond.

14. BERENDS, J.; CARRARA, W.; VOLLERS, H. *Analytical Report 5: Barriers in working with Open Data*. Luxembourg: Publications Office of the European Union, 2020. ISBN 978-92-78-41922-6. Available also from: https://data.europa.eu/sites/default/files/edp_analytical_report_n5_-_barriers_in_open_data.pdf.

15. CARRARA, W.; SAN CHAN, W.; FISCHER, S.; STEENBERGEN, E. van. *Creating Value through Open Data: Study on the Impact of Re-use of Public Data Resource*. Luxembourg: Publications Office of the European Union, 2015. ISBN 978-92-79-52791-3. Available also from: https://data.europa.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf.

16. CARRARA, W.; FISCHER, S.; STEENBERGEN, E. van. *Analytical Report 2: E-skills and Open Data*. Luxembourg: Publications Office of the European Union, 2020. ISBN 978-92-78-41887-8. Available also from: https://data.europa.eu/sites/default/files/edp_analytical_report_n2_-_e-skills.pdf.

17. SADIQ, S.; INDULSKA, M. Open data: Quality over quantity. *International Journal of Information Management*. 2017, vol. 37, no. 3, pp. 150–154. ISSN 0268-4012. Available from DOI: https://doi.org/10.1016/j.ijinfomgt.2017.01.003.

18. THEREAUX, O. *Data and Covid-19: why standards matter*. Open Data Institute, 2020-06-15. Available also from: https://theodi.org/article/data-and-covid-19-why-standards-matter/.

19. AROCENA, P.; GLAVIC, B.; MECCA, G. et al. Benchmarking Data Curation Systems. *IEEE Data Engineering Bulletin*. 2016, vol. 39, pp. 47–62. Available also from: https://www.researchgate.net/publication/305333003_Benchmarking_Data_Curation_Systems.

20. EU, European Data Portal. *What makes quality open data*. Publications Office of the European Union. Available also from: https://data.europa.eu/elearning/en/module5/#/id/co-01.

21. *5 Star Linked Data*. World Wide Web Consortium. Available also from: https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data.

22. *What is an open data certificate?* Salsa. Available also from: https://salsadigital.com.au/insights/what-is-open-data-certificate.

23. NEVES, F. T.; DE CASTRO NETO, M.; APARICIO, M. The impacts of open data initiatives on smart cities: A framework for evaluation and monitoring. *Cities*. 2020, vol. 106, p. 102860. ISSN 0264-2751. Available from DOI: https://doi.org/10.1016/j.cities.2020.102860.

24. VÍTA, M. *Vybrané techniky analýzy datové kvality a jejich aplikace na data z Centrální evidence projektů VaVaI*. Ministerstvo vnitra České republiky, 2021-10-29. Available also from: https://data.gov.cz/%C4%8Dl%C3%A1nky/techniky-anal%C3%BDzy-datov%C3%A9-kvality-cep.

25. MALINA, J. *Otevřená data v České republice – aktuální stav*. Otevřená data 2021, Praha. Ministerstvo vnitra České republiky, 2021-11-05. Available also from: https://docs.google.com/presentation/d/14XTs3ugaGXX7iapxzZskQLnt-OdVLOOi/view#slide=id.p13.

26. KLÍMEK, J. *Příklady špatné praxe v oblasti otevřených dat*. Ministerstvo vnitra České republiky, 2021-01-19. Available also from: https://opendata.gov.cz/%C5%A1patn%C3%A1-praxe:start.

27. EU, European Commission. *Summary report of the public consultation on the European strategy for data*. European Commission, 2020-07-24. Available also from: https://ec.europa.eu/digital-single-market/en/news/summary-report-public-consultation-european-strategy-data.

28. KLÍMEK, J. *Špatné dělení dat do distribucí datové sady*. Ministerstvo vnitra České republiky, 2020-12-10. Available also from: https://opendata.gov.cz/%C5%A1patn%C3%A1-praxe:%C5%A1patn%C3%A9-d%C4%9Blen%C3%AD-distribuc%C3%AD.

29. *DCAT Application Profile for data portals in Europe: Specification describing public sector datasets*. European Commission. Available also from: https://ec.europa.eu/isa2/solutions/dcat-application-profile-data-portals-europe_en.

30. KLÍMEK, J. *Rozhraní katalogů otevřených dat: DCAT-AP-CZ*. Ministerstvo vnitra České republiky, 2021-01-11. Available also from: https://ofn.gov.cz/rozhran%C3%AD-katalog%C5%AF-otev%C5%99en%C3%BDch-dat/2021-01-11/.

31. EU, European Data Portal. *Choosing the right format for open data*. Publications Office of the European Union. Available also from: https://data.europa.eu/elearning/en/module9/#/id/co-01.

32. KLÍMEK, J. *Technické standardy pro datové sady na stupni otevřenosti 3*. Ministerstvo vnitra České republiky, 2020-09-02. Available also from: https://opendata.gov.cz/standardy:technicke-standardy-pro-datove-sady-na-stupni-3.

33. GUPTA, L. *JSON Schema*. REST API Tutorial. Available also from: https://restfulapi.net/json-schema/.

34. RICHER, J. *Open Data Quality, Standardization and Why we Need a Schema Catalog*. Frictionless Data, Open Knowledge Foundation, 2020-04-23. Available also from: https://frictionlessdata.io/blog/2020/04/23/table-schema-catalog/#schemas-as-standards.

35. *Otevřené formální normy (OFN)*. Ministerstvo vnitra České republiky. Available also from: https://data.gov.cz/ofn/.

36. KLÍMEK, J. *Otevřená data a otevřené formální normy*. Ministerstvo vnitra České republiky, 2020-10-02. Available also from: https://data.gov.cz/%C4%8Dl%C3%A1nky/otev%C5%99en%C3%A9-form%C3%A1ln%C3%AD-normy-01-%C3%BAvod.

37. *Jak mám jako poskytovatel použít otevřené formální normy*. Ministerstvo vnitra České republiky. Available also from: https://data.gov.cz/ofn/pou%C5%BEit%C3%AD-poskytovateli/.

38. ŠKOP, M. *Z otevřených dat rovnou na plakát*. Ministerstvo vnitra České republiky, 2021-05-04. Available also from: https://data.gov.cz/%C4%8Dl%C3%A1nky/z-otev%C5%99en%C3%BDch-dat-rovnou-na-plak%C3%A1t.

39. WANG, R. Y.; STRONG, D. M. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*. 1996, vol. 12, pp. 5–33. Available also from: http://mitiq.mit.edu/Documents/Publications/TDQMpub/14_Beyond_Accuracy.pdf.

40. DVOŘÁK, M.; SPÁL, M.; MAREK, J.; KLÍMEK, J. *Turistické cíle: Doporučení 1. července 2020*. Ministerstvo vnitra České republiky, 2020-07-01. Available also from: https://ofn.gov.cz/turistick%C3%A9-c%C3%ADle/2020-07-01/.

41. ABEDJAN, Z. Data Profiling. In: *Encyclopedia of Big Data Technologies*. Ed. by SAKR, S.; ZOMAYA, A. Springer International Publishing, 2018, pp. 1–6. ISBN 978-3-319-63962-8. Available from DOI: 10.1007/978-3-319-63962-8_8-1.

42. HEER, J.; HELLERSTEIN, J. M.; KANDEL, S. Data Wrangling. In: *Encyclopedia of Big Data Technologies*. Ed. by SAKR, S.; ZOMAYA, A. Y. Springer International Publishing, 2019, pp. 584–591. ISBN 978-3-319-77525-8. Available from DOI: 10.1007/978-3-319-77525-8_9.

43. SPORNY, M.; LONGLEY, D. et al. *JSON-LD 1.1: A JSON-based Serialization for Linked Data*. World Wide Web Consortium, 2020-07-16. Available also from: https://www.w3.org/TR/json-ld11/#the-context.

44. GOESSNER, S. *Converting Between XML and JSON*. O'Reilly Media, 2006-05-31. Available also from: https://www.xml.com/pub/a/2006/05/31/converting-between-xml-and-json.html.

45. GARCIA, J.; MOLINA, J. M.; BERLANGA, A.; PATRICIO, M. A. Data Fusion. In: *Encyclopedia of Big Data Technologies*. Ed. by SAKR, S.; ZOMAYA, A. Cham: Springer International Publishing, 2018, pp. 1–6. ISBN 978-3-319-63962-8. Available from DOI: 10.1007/978-3-319-63962-8_5-1.

46. CODD, E. F. Missing Information (Applicable and Inapplicable) in Relational Databases. *SIGMOD Rec.* 1986, vol. 15, no. 4, p. 53. ISSN 0163-5808. Available also from: https://doi.org/10.1145/16301.16303.

A Complete JSON schema

Below is the complete JSON schema, which was used for the creation of the proposed evaluation model. The schema is taken from the official GitHub repository[1] intended for OFNs development.

```
{
    "$schema": "https://json-schema.org/draft/2019-09/schema",
    "$id": "https://ofn.gov.cz/turistické-cíle/2020-07-01/
        schémata/turistický-cíl.json",
    "type": "object",
    "title": "Turistický cíl",
    "required": ["typ"],
    "allOf": [{
        "$ref": "https://ofn.gov.cz/věc/2020-07-01/
            schémata/věc.json"
    }, {
        "properties": {
            "typ": {
                "oneOf": [{
                    "const": "Turistický cíl"
                }, {
                    "type": "array",
                    "contains": {
                        "const": "Turistický cíl"
                    },
                    "items": {
                        "type": "string"
                    }
                }]
            },
            "kouření_povoleno": {
                "type": "boolean",
                "title": "Kouření povoleno",
```

1. https://github.com/opendata-mvcr/otevrene-formalni-normy/blob/master/turistické-c%C3%ADle/2020-07-01/schémata/turistický-c%C3%ADl.json

```
        "examples": [ true, false ]
    },
    "veřejná_přístupnost": {
        "type": "boolean",
        "title": "Veřejná přístupnost",
        "examples": [ true, false ]
    },
    "kapacita": {
        "$ref": "https://ofn.gov.cz/základní-datové
            -typy/2020-07-01/schémata/množství.json"
    },
    "vlastník": {
        "type": "array",
        "items": {
            "$ref": "https://ofn.gov.cz/lidé-a-osoby/
                2020-07-01/schémata/člověk-či-osoba.json"
        }
    },
    "provozovatel": {
        "$ref": "https://ofn.gov.cz/lidé-a-osoby/
            2020-07-01/schémata/člověk-či-osoba.json"
    },
    "umístění": {
        "$ref": "https://ofn.gov.cz/umístění/
            2020-07-01/schémata/umístění.json"
    },
    "kontakt": {
        "type": "array",
        "items": {
            "$ref": "https://ofn.gov.cz/kontakty/
                2020-07-01/schémata/kontakt.json"
        }
    },
    "bezbariérovost": {
        "$ref": "https://ofn.gov.cz/bezbariérovost/
            2020-07-01/schémata/bezbariérovost.json"
    },
    "vstupné": {
```

```
                "$ref": "https://ofn.gov.cz/vstupné/
                    2020-07-01/schémata/více-vstupných.json"
            },
            "otevírací_doba": {
                "$ref": "https://ofn.gov.cz/časová
                    -specifikace/2020-07-01/schémata/seznam
                    -časových-specifikací.json"
            },
            "typ_turistického_cíle": {
                "type": "string",
                "format": "iri",
                "pattern": "^https\\:\/\/data\\.mvcr\\.gov\\
                      .cz\/zdroj\/číselníky\/typy-turistických-
                      cílů\/položky\/.*$",
                "title": "Typ turistického cíle",
                "examples": [
                    "https://data.mvcr.gov.cz/zdroj/číselníky/
                    typy-turistických-cílů/položky/přírodní"
                ]
            },
            "jazyk": {
                "type": "array",
                "items": {
                    "type": "string",
                    "format": "iri",
                    "pattern": "^http\\:\/\/publications\\
                        .europa\\.eu\/resource\/authority\
                        /language\/.*$",
                    "title": "Jazyk",
                    "examples": [
                        "http://publications.europa.eu/
                        resource/authority/language/BOS"
                    ]
                }
            }
        }
    }]
}
```

B Recommendations for examined datasets

Because each dataset has its own mix of strengths and shortcomings, we provide a list of custom recommendations for each of them to increase their data quality. The list is not complete; it contains advice to remedy only the most severe deficiencies observed during the evaluation.

Brno

1. Model objects as tourist points of interest, not solely as geographical points.
2. Change features names to match the naming convention defined by the standard. Use the Czech language for feature names.
3. Fill in the information about opening hours, entrance fee, and contact for individual objects instead of providing null values.
4. Remove duplicates regarding the objects' location, The same information is provided within four different features.
5. Add missing *@context* and *typ* features.

Děčín

1. Change the file format because XLSX as a proprietary file format contradicts the idea of open data. Change data structure from tabular to hierarchical.
2. Change features names to match the naming convention defined by the standard. Use the Czech language for feature names.
3. Add missing features that are required, especially *vstupné* and *otevírací doba*.
4. Remove duplicate features.
5. Remove features that do not contain any values.
6. Unify data types used within individual features.
7. Fill in missing values.

Hradec Králové

1. Change features names to match the naming convention defined by the standard. Use the Czech language for feature names.
2. Model objects as tourist points of interest, not solely as geographical points.
3. Add missing required features, including *vstupné*, *otevírací doba*, *type* and *@context*.
4. Remove duplicate information about objects' locations.

Huntířov

1. Change file format from tabular to hierarchical.
2. Change features names to match the naming convention defined by the standard.
3. Add missing required features, in particular *iri*, *vstupné*, *otevírací doba*, *typ* and *@context*.
4. Fill in missing values.

Ostrava

1. Change file encoding from Western (Windows 1252) to UTF-8.
2. Change file format from tabular to hierarchical.
3. Change features names to match the naming convention defined by the standard.
4. Add missing required features such as *typ* and *@context*.
5. Remove duplicate information about objects' locations.
6. Fill in missing values, especially the objects' names.

Praha

1. Model objects as tourist points of interest, not solely as geographical points.
2. Change features names to match the naming convention defined by the standard. Use the Czech language for feature names.
3. Add missing required features, including *iri*, *vstupné*, *otevírací doba*, *typ* and *@context*.