



Assignment of bachelor's thesis

Title:	Official notice boards
Student:	Jakub Kučera
Supervisor:	Mgr. Martin Mareš
Study program:	Informatics
Branch / specialization:	Knowledge Engineering
Department:	Department of Applied Mathematics
Validity:	until the end of summer semester 2022/2023

Instructions

Public data are a popular topic nowadays. The amendment to the Act No. 106/1999 forces public authorities to publish official notice boards on-line. The main goal of this thesis is to analyse these new open data and present them. The parts of the thesis are:

- Obtain a list of municipalities and links to their official notice boards (primarily focus on sources from data.gov.com).
- Analyse the data from official notice boards with regards to the amendment of Section 3 paragraph 9 Act No. 106/1999 (checks compliance with the law, displays the characteristics of the inserted document and offers basic analysis).
- Design and implement a web application that shows your results and enables some form of search within official notice boards.
- The resulting web application has to be intuitive and understandable by the general public.

Electronically approved by Ing. Karel Klouda, Ph.D. on 20 January 2022 in Prague.

Bachelor's thesis

OFFICIAL NOTICE BOARDS

Jakub Kučera

Faculty of Information Technology
Department of Applied Mathematics
Supervisor: Mgr. Martin Mareš
May 12, 2022

Czech Technical University in Prague
Faculty of Information Technology

© 2022 Jakub Kučera. Citation of this thesis.

This thesis is school work as defined by Copyright Act of the Czech Republic. It has been submitted at Czech Technical University in Prague, Faculty of Information Technology. The thesis is protected by the Copyright Act and its usage without author's permission is prohibited (with exceptions defined by the Copyright Act).

Citation of this thesis: Kučera Jakub. *Official notice boards*. Bachelor's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2022.

Contents

Acknowledgments	vi
Declaration	vii
Abstrakt	viii
List of abbreviations	ix
Introduction	1
1 Analysis	3
1.1 Official notice board	3
1.2 Act on Free Access to Information	3
1.3 Open formal norms for official notice boards	3
1.3.1 Official notice board metadata	4
1.3.2 Official notice board data	4
1.4 Existing projects	4
1.4.1 Edesky	4
1.4.2 Open Formal Norms	4
1.4.3 Open Data Portal	6
1.4.4 Notice boards	6
1.5 Data sources	6
1.5.1 National Open Data Catalogue	6
1.5.2 Statistical metainformation system	6
1.5.3 Wikidata	6
1.5.4 Edesky	7
1.5.5 individual websites of data providers	7
1.6 SPARQL	7
2 Design	9
2.1 Data importer	9
2.1.1 import_all_data	9
2.1.2 import_new_data	10
2.1.3 Repeated run of the data importer	10
2.2 Database	10
2.2.1 Municipality	10
2.2.2 OfficialNoticeBoard	12
2.2.3 Notice	12
2.2.4 NoticeDocument	12
2.3 Website	12
2.3.1 About page	12
2.3.2 Statistics	12
2.3.3 List of municipalities	14
2.3.4 Detailed municipality view	14

2.3.5	List of official notice boards	15
2.3.6	Detailed official notice board view	15
2.3.7	List of document	15
2.4	Statistics	17
3	Implementation	23
3.1	Used technologies	23
3.2	System structure	23
3.3	Data importer	23
3.3.1	Create database tables	24
3.3.2	Get the list of municipalities and municipality parts	24
3.3.3	Map RUIAN to IČO	24
3.3.4	Mark municipality offices with extended competence	25
3.3.5	Get the list of available official notice boards metadata	25
3.3.6	Download boards + extract data	27
3.3.7	Download documents + extract text	27
3.3.8	Modifications for <code>import_new_data</code>	27
3.4	How to install and run	27
3.5	Possible future improvements	28
4	Summary	29
	Contents of the enclosed media	33

List of Figures

1.1	Example of SPARQL Triple	7
2.1	The 3 parts that the system is divided into.	9
2.2	Database model diagram	11
2.3	Illustration of the About page	13
2.4	Illustration of the Statistics page	13
2.5	Illustration of the List of municipalities page	14
2.6	Illustration of the Detailed municipality part page for Prague 12	15
2.7	Illustration of the List of official notice boards page	16
2.8	Illustration of the Detailed official notice board page	16
2.9	Illustration of the List of documents page	17
2.10	How many municipality offices with extended competence publish boards	18
2.11	Out of all boards published by municipality offices, how many were published by the ones with extended competence	19
2.12	Number of errors by municipalities that violate the minimum specification	19
2.13	Municipalities with the most amount of unreachable download URLs for documents	20
2.14	Document file extensions	20
2.15	How many of the PDFs contain a text layer and how many do not	21
2.16	What municipality offices post the most PDFs without a text layer	21

List of code listings

1.1	Example of minimal official format specification for official notice board	5
3.1	Example of a list of municipalities in JSON downloaded from NKOD.	25
3.2	SPARQL query, that will return all municipality offices with extended competence	25
3.3	JSON response with a list of municipality offices with extended competence	26
3.4	Example of a simple SPARQL query, that will return download URLs to all published official notice boards.	26

I'd like to thank my thesis supervisor Mgr. Martin Mareš and also Ing. Marek Sušický and Mgr. Adam Szabó, for all of their help and many useful advises, while I was working on this thesis.

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis. I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended. In accordance with Article 46(6) of the Act, I hereby grant a nonexclusive authorization (license) to utilize this thesis, including any and all computer programs incorporated therein or attached thereto and all corresponding documentation (hereinafter collectively referred to as the “Work”), to any and all persons that wish to utilize the Work. Such persons are entitled to use the Work in any way (including for-profit purposes) that does not detract from its value. This authorization is not limited in terms of time, location and quantity. However, all persons that makes use of the above license shall be obliged to grant a license at least in the same scope as defined above with respect to each and every work that is created (wholly or in part) based on the Work, by modifying the Work, by combining the Work with another work, by including the Work in a collection of works or by adapting the Work (including translation), and at the same time make available the source code of such work at least in a way and scope that are comparable to the way and scope in which the source code of the Work is made available.

In Prague on May 12, 2022

.....

Abstract

The main goal of this thesis is to analyze published official notice boards in form of open data and present the results on the web. In order to analyze the compliance of the published official notice boards, a list of municipalities is obtained. In the next step, a list of links to all published official notice boards is obtained. These boards then get downloaded and have data extracted from them. If some of these boards have links to attached documents, those will get downloaded as well and text will be extracted from them. All of these processed data will be stored in a database. This data can be used later on. For example on the created website, where a user can look through the process data, or at one of many charts, showing statistics of the data.

Keywords open data, Python, official notice boards, data analysis, municipality offices, web application

Abstrakt

Hlavní cíl této práce, je analyzovat webové úřední desky, které byly publikovány jako otevřená data, a následně výsledky zobrazit na webové stránce. K tomu, aby se mohly tyto publikované desky analyzovat, je nejdříve získán seznam všech obcí. V dalším kroku se stáhne seznam odkazů na všechny publikované úřední desky. Soubory těchto desek se následně stáhnou a jsou z nich extrahována data. V případě, že některá z těchto desek obsahuje odkaz na příložený dokument, tak se tento dokument stáhne a extrahuje se z něj text. Všechna tato zpracovaná data se uloží do databáze, aby se mohla později použít. Například na vytvořené webové stránce si je uživatel může prohlédnout, nebo si může prohlédnout jeden z mnoha grafů, které zobrazují jejich různé statistiky.

Klíčová slova otevřená data, Python, úřední desky, analýza dat, obecní úřady, webová aplikace

List of abbreviations

API	Application Programming Interface
CSS	Cascading Style Sheets
HTML	HyperText Markup Language
IČO	Identifikační Číslo Osoby (Person Identified Number)
IRI	Internationalized Resource Identifier
JSON	JavaScript Object Notation
NKOD	Národní Katalog Otevřených Dat (National Open Data Catalogue)
OCR	Optical Character Recognition
OFN	Open Formal Norms
ORM	Object-Relational Mapping
PDF	Portable Document Format
REST	REpresentational State Transfer
RDF	Resource Description Framework
RÚIAN	Registr Územní Identifikace, Adres a Nemovitostí
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
URI	Uniform Resource Identifier
URL	Uniform Resource Locator

Introduction

From the 1st of February 2022, Act No. 261/2021 has come into effect. This Act is an amendment to the Act No. 106/1999, which now among other states that state administration authorities, regional offices, and municipality offices with extended competence have to publish their official notice boards in form of open data [1]. The act was followed by a recommended format specification that the publisher of official notice boards should follow. Some of these publishers however do follow this specification only to some extent or they do not publish anything at all.

The resulting web application can be useful for people who are looking to use the open data of official notice boards and who want to know the level of adaptation of the specification. This information could help others decide whether they can be used for their own use cases. It can also be used by people who want to at look the official notice boards from a single website, or by people who wish to search through text amongst all of the documents that are linked from all the official notice boards. The increased transparency of the offices' lack of compliance with the aforementioned act might motivate them to improve it.


I chose this topic since I was interested in how well the municipality and other offices follow the official notice board specification or if they publish them at all. I also wanted to learn more about open data here in the Czech Republic and about the National Open Data Catalog.

The first goal is to obtain a list of municipalities and links to their official notice boards. The next goal is to analyze the data from the official notice boards with regards to the amendment of Section 3 paragraph 9 of Act No. 106/1999, whereby analysis is meant checking compliance with the mentioned Act, displaying characteristics of the inserted documents, and offering basic analysis. After that, the task is to design and implement a web application that shows the results of the data analysis and enables some form of search within official notice boards. The last goal is that the resulting web application has to be intuitive and understandable by the general public.

The first part starts by focusing on the legal act in question, and what it exactly means in regards to the official notice boards and their format. Then it also takes a look at the existing options and possible data sources.

The second part starts with the high-level design of how the system that obtains data and processes them, will work. The next section is about the database and its' models. Then, it describes the layout of each page on the website and its use cases. The last part is more about the graphs, representing the various data characteristics.

The last and the longest part is the third one, which is about the implementation. It begins with getting all the data and processing it. Then it moves on to the website part.



Chapter 1

Analysis

1.1 Official notice board

The official notice board is a publicly accessible area intended for the publication of legal regulations, decisions, and other documents of administrative bodies and courts. It must be publicly accessible at all times and it must be kept in two forms at the same time, in a physical and an electronic form [2].

1.2 Act on Free Access to Information

The Act No. 261/2021, which came into effect on the 1st of February 2022, amends the Act No. 106/1999 on Free Access to Information. This amendment states in the Section 5a paragraph 3 that state administration authorities, regional offices, and municipality offices with extended competence have to publish metadata of information (that information that was published in a way allowing remote access) and metadata of these boards in form of open data.

But what exactly are open data? There are multiple definitions, but from the legal point of view, according to the Act No. 106/1999 the Section 3, the open data are any data that are machine-readable, are in an open format, and are published in a way allowing remote access.

The machine-readable format is understood as a format of data with such structure, that it allows any software easily find, recognize and obtain certain information, including separate details and their inner structure.

The open format is defined as a format of data, which is not dependent on specific hardware or software and could be accessible by the general public without any restrictions, which would prevent them the use of the information contained in the data file.

The section 4b of the Act No. 106/1999 also states that metadata and data format of published data should match the open formal norms (OFNs) as close as possible .

Open Formal Norms are specifications for the syntactic and semantic form of datasets published as open data that follow modern web standards [3].

1.3 Open formal norms for official notice boards

In this section, I write about the recommended specification defined by the OFNs that the published official notice boards should comply with. Not only that there is a specification for boards' data, but there is also a specification for the boards' metadata.

1.3.1 Official notice board metadata

The most important attribute of the boards' metadata is the link to the specification of the boards' data, alongside the link with the current version included. Another important attribute to specify is the JSON Schema in the case of JSON-LD distribution of boards' data. This schema is talked about in right in the next subsection. By specifying the board specification, it enables this board to be found, when listing all of the available boards. This is actually how the already mentioned application on ofn.gov.cz gets the list of all boards [4].

1.3.2 Official notice board data

The official specification of board data includes many attributes, many of them are however only optional, "depending on what the systems of data providers are capable of". But there is still a "minimal scope" that needs to be complied with, an example of it is at 1.1. The first attribute is the IRI of the entire board, which acts as an identifier, then there is the URL to the website with the electronic version of the board (not the URL, from which the board has been downloaded from). The next attribute is the publisher of the board (it is not just the name of the publisher, but a link to an entry about them). The last mandatory attribute of the boards directly is a list of attached information on the board. It is not against the specification if this list is empty, however, it would be extremely improbable since there's probably going to be at least something on a board, that was published by somebody.

The attached information has some mandatory attributes, which are: IRI, URL, name, published date, and date of relevance.

There are many optional attributes. One of which worth mentioning is for example a list of documents related to single information. Each of these documents can have a URL, which should not lead to a website showing information about it and potentially allowing the user to download it, like the other mentioned URLs. Instead, it should be a direct download link, which would make it much easier to download using automated software. If however, a publisher is not satisfied with the available optional attributes, then they can define their own attributes.

1.4 Existing projects

In this section, I write about various other projects that are related to this one.

1.4.1 Edesky

Edesky is a web portal with the main goal of showing documents that are part of official notice boards published electronically, but not necessarily in form of open data. Since this portal is closed-source, it is not known how the documents are obtained, however, it is probably directly from the websites of their publishers. Some of its other features are searching and sorting documents, sending emails about newly published documents, API access, and recently added full-text search on documents [5].

1.4.2 Open Formal Norms

Directly on the website for OFNs ofn.gov.cz, which is also where the OFN for official notice boards is defined, there is a simple application, which retrieves all of the published boards, that are available on the NKOD – National Open Data Catalogue. This application then lists these boards, along with a list of linked information. However, if there is a problem with the board's data, like an invalid link, or incompatible data with the specification, it should show an appropriate error describing it [6].


```
{
  "@context":
    ↪ "https://ofn.gov.cz/úřední-desky/2021-07-20/kontexty/úřední-deska.jsonld",
  "typ": "Úřední deska",
  "iri": "https://data.mojeobec.cz/zdroj/úřední-deska",
  "stránka": "https://web.mojeobec.cz/úřední_deska/",
  "provozovatel": {
    "typ": "Osoba",
    "ičo": "00258245"
  },
  "informace": [{
    "typ": ["Digitální objekt", "Informace na úřední desce"],
    "iri":
      ↪ "https://data.mojeobec.cz/zdroj/úřední-deska/informace/2018-13",
    "url": "https://web.mojeobec.cz/úřední_deska/2018-13",
    "název": {
      "cs": "Podpora spolkového života ve městě"
    },
    "vyvěšení": {
      "typ": "Časový okamžik",
      "datum": "2018-01-20"
    },
    "relevantní_do": {
      "typ": "Časový okamžik",
      "datum": "2019-02-20"
    }
  }
}]
}
```

■ **Code listing 1.1** Example of minimal official format specification for official notice board

1.4.3 Open Data Portal

Directly on the Open Data Portal, which is also from where NKOD can be accessed, there are some extra pages showing various analytics, however, those include all datasets available and not only official notice boards [7].

1.4.4 Notice boards

There is already an existing project that belongs under OpenDataLabCZ, which is called "notice boards". It can download documents from municipality offices, per users' choice. Since this project was developed in 2019, there was no law specifying anything regarding publishing official notice boards as open data. Because of that, this project attempts to get the boards' data directly from the websites of municipality offices. [8].

1.5 Data sources

This section is about the various data sources that are used later on. It is about their pros and cons for different use cases. Then there are also some alternative data sources that are not used, but I believe that it is worth mentioning them.

1.5.1 National Open Data Catalogue

"The National Open Data Catalogue (NKOD) is the public administration information system that is accessible in a manner allowing remote access used for recording information published as open data." [9]

People can search for data directly on the NKOD website, however there it also provides a SPARQL endpoint.

1.5.2 Statistical metainformation system

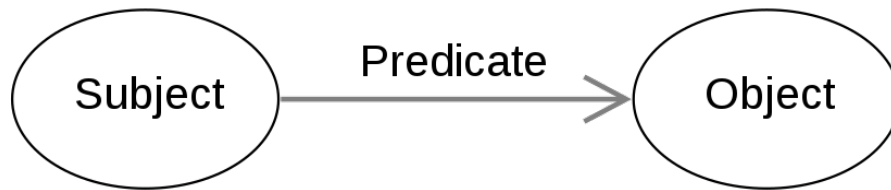
It is a system made by the Czech Statistical Office and as the name suggests, you can find there various metainformation datasets.

Metainformation is knowledge of objects described by statistical metadata. Statistical metadata comprise data and other documentation that describe objects in a formalised way. Statistical metadata provide information on data and about processes of producing and using data. Statistical metadata describe statistical data and - to some extent - processes and tools involved in the production and usage of statistical data. [10]

The data there are divided into three groups: code lists, classifications, and indicators. Some of the relevant code lists are for example list of all municipalities and a list of all municipality parts [10]. Statistical metainformation system was not used, since all of the relevant datasets are also available on the NKOD, and it does not provide an API like NKOD does.

1.5.3 Wikidata

Wikidata is secondary storage for structured data, that is used by wikis of the Wikimedia movement. That means that pages on Wikipedia can use data directly from Wikidata. A great advantage of Wikidata is its SPARQL endpoint, which allows people (who know how to work with it) to easily query information [11]. While there are no official notice boards stored on Wikidata, there is information about their publishers, like the municipality offices.



■ **Figure 1.1** Example of SPARQL Triple

1.5.4 Edesky

As mentioned in the Existing projects section 1.4.1, Edesky provides API access, that allows the listing of official notice boards and also searching of documents published on said boards.

1.5.5 individual websites of data providers

As mentioned in part about the existing notice board project 1.4.4, even the website of each municipality office can be used as a data source. However, this is not ideal, for the use of automated software, since some of these websites can vary a lot, so the software would have to be modified a lot so that it can work correctly with all of the websites.

1.6 SPARQL

SPARQL is a query language for graph databases. The querying works using the RDF triplets. As can be seen in 1.1, a triplet consists of a subject, a predicate, and an object [12]. An example of a triplet can be "Official notice board" "published.by" "municipality office of Prague 1".

Chapter 2

Design

In this chapter, I talk about the higher level design of the entire system, without any implementation details. That includes how the script that imports and processes all of the data works, database models, the layout of separate pages of the website, and statistics charts.

Even though it was not mentioned in any of the goals, I designed the system in a way, that it should be easy to deploy in production. By that, I mean that most of the processes are automated and don't require much interaction from the person who maintains the system.

The system is separated into 3 main parts 2.1:

- Data importer – Retrieves all necessary data from external sources and processes them.
- Database – Stores the processed data.
- Website – Presents processed data to a user.

2.1 Data importer

Since we want to analyze the data of official notice boards, we have to have them in the first place and possibly also process them somehow. That's what the data importer is for.

The *data importer* has two commands:

- `import_all_data`
- `import_new_data`

2.1.1 `import_all_data`

This command will create tables in database and import all of the necessary data, including list of municipalities, published official notice boards, and the documents that are attached to the boards.



■ **Figure 2.1** The 3 parts that the system is divided into.

2.1.2 import_new_data

This command will only import newly published official notice boards. If an existing board has been deleted on NKOD, it will also be deleted by data importer. If the existing board is still accessible on NKOD, *data importer* will download it and check if there are any new information on it, which it will also import.

If `import_new_data` is called, even though the `import_all_data` has not yet been called, it will be run instead.

2.1.3 Repeated run of the data importer

Since there is always new information being posted on official notice boards, the data importer needs to be run more than once. However running all of the steps all over again, would be inefficient and time-consuming.

2.2 Database

All of the data that get downloaded and processed by data importer need to be stored somehow so that it can be later accessed by the website.

To enable full-text search on extracted text from documents, there could be another part of the system, which would be Elasticsearch [13]. However to prevent unnecessary complexity, it can be replaced by a database engine, that does support full-text search. One such database engine is PostgreSQL [14].

Database models, that are used to store data:

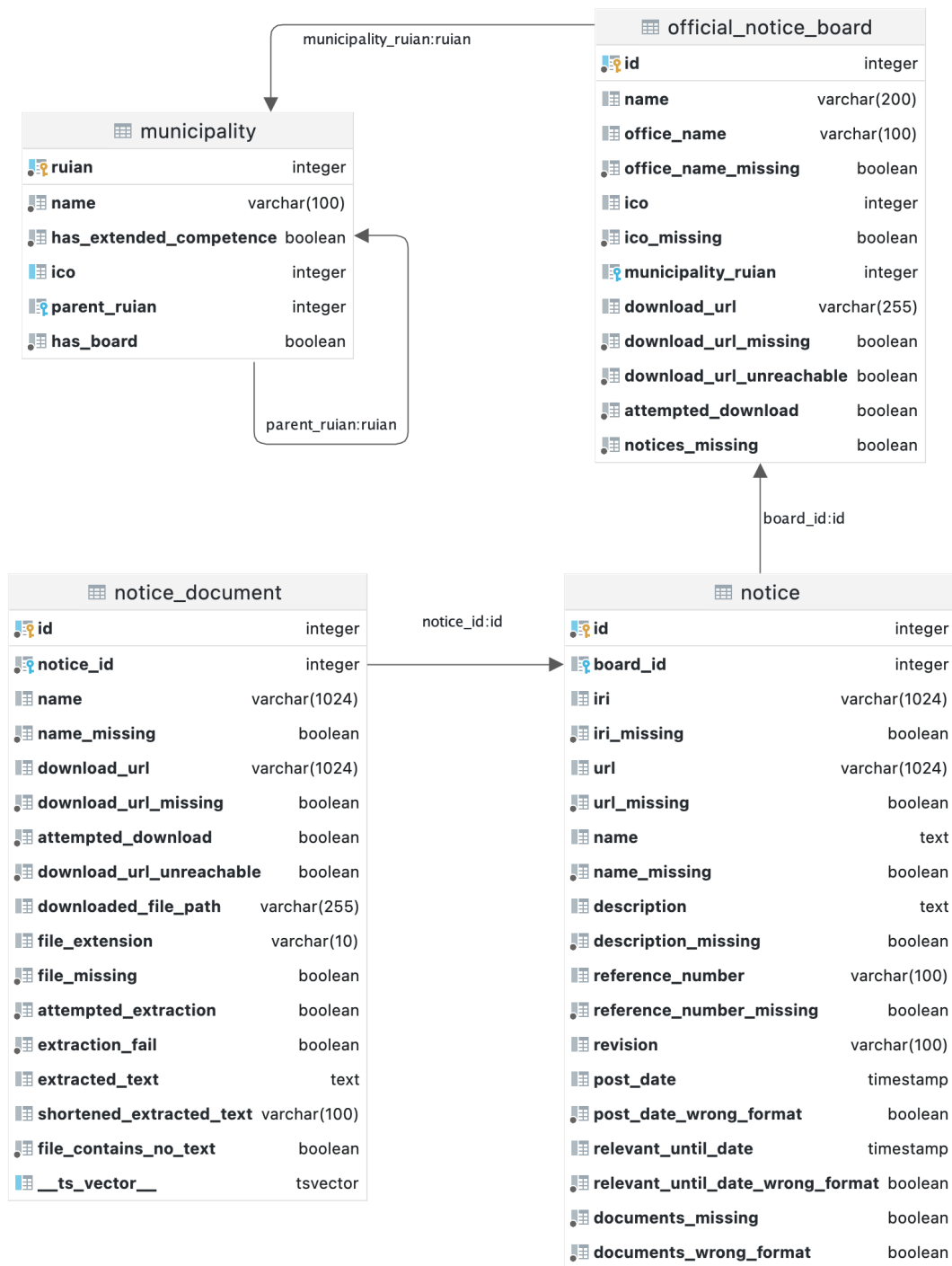
- Municipality
- OfficialNoticeBoard
- Notice
- NoticeDocument

The relationships between the models and overview of all of their attributes is at 2.2.

2.2.1 Municipality

Municipality represents a single municipality or municipality part. It has an attribute `has_extended_competence`, which represents, whether the municipality office, that belongs under the municipality has extended competence. For identification, it uses the RUIAN of the municipality/municipality part. While `IČO` is one of the attributes, it is not used as an identifier, since some municipality parts don't have their own `IČO`, so they use the one of their superior office, which means, that is not unique.

It might be a bit unusual choice for *Municipality* model to store municipality part as well. A different approach might be to have separate models for both municipalities and municipality parts (*Municipality*, *MunicipalityPart*). Another, more complex option would be to have a more general model for public authorities (*PublicAuthority*) and then models like *Municipality* and *MunicipalityPart*, which would be a specialization of *PublicAuthority*. These alternate options were not used, since they would be unnecessary for its use case and would add unnecessary complexity.



■ Figure 2.2 Database model diagram

2.2.2 OfficialNoticeBoard

OfficialNoticeBoard represent a single official notice board. It includes all of the attributes from the minimal scope specification 1.3.2.

2.2.3 Notice

Notice represents a single piece of information on an official notice board. It is named *Notice*, to signify the relation to *OfficialNoticeBoard*. Like *OfficialNoticeBoard*, it also includes all of the attributes from the minimal scope specification 1.3.2 and some optional ones as well, like description, reference_number, revision, and a list of attached documents.

2.2.4 NoticeDocument

NoticeDocument represents a document that is attached to a single information. Not only does it represent the metadata of the document, that can be found on the official notice board, but it also represents the actual document file, so it can store the text that gets extracted from it.

2.3 Website

In this part, I write about the general layout of the website and its individual pages.

List of types of pages:

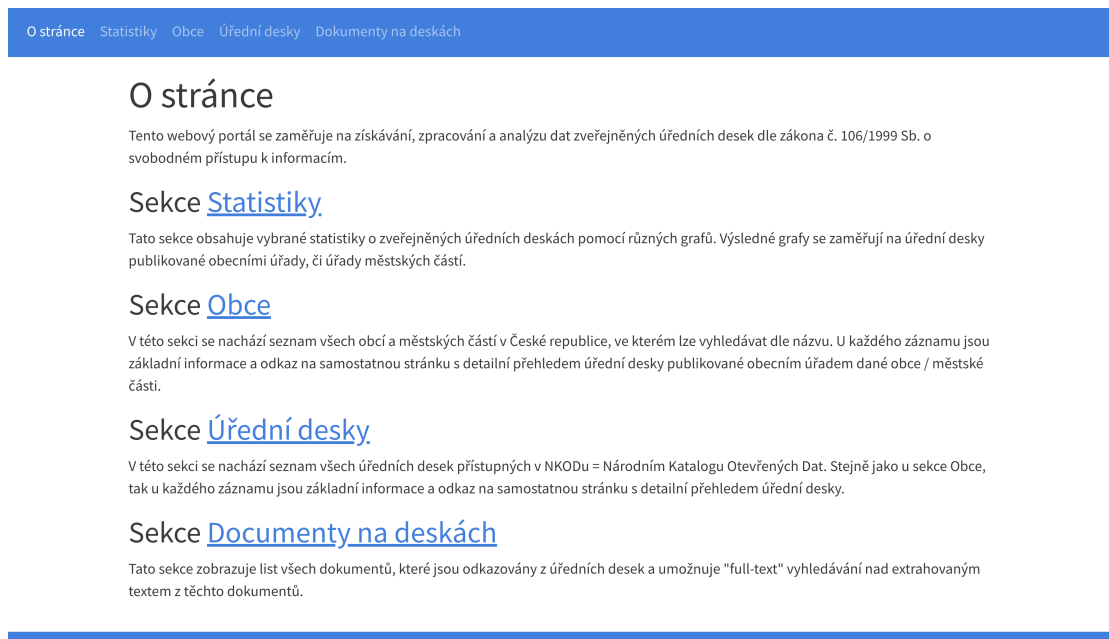
- About page
- Statistics
- List of municipalities
- Detailed view of a specific municipality
- List of official notice boards
- Detailed view of a specific official notice board
- List of documents

2.3.1 About page

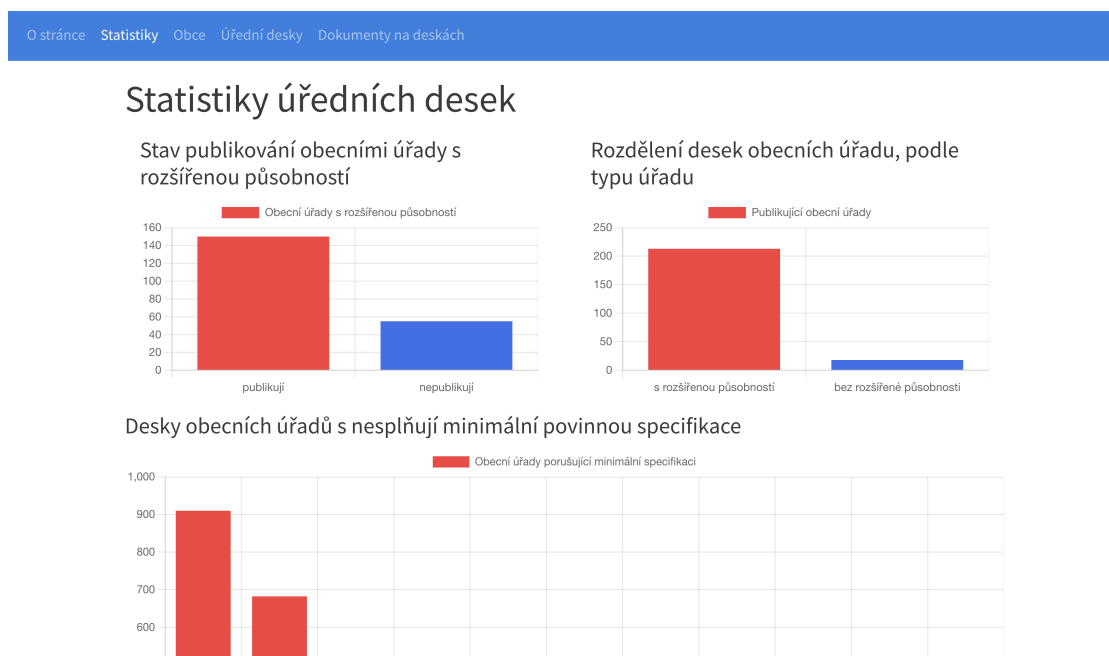
This page informs the users, about the whole website, and what can be found on each page. An illustration of the **About page** can be found at 2.3

2.3.2 Statistics

This page contains chosen statistics about the published official notice boards using various charts, an illustrated example is at 2.4. Most of the charts focus on boards that were published by municipality offices and municipality part offices. More about the statistics is in 2.4.



■ Figure 2.3 Illustration of the About page



■ Figure 2.4 Illustration of the Statistics page

Obce

Název	RÚIAN	IČO	Úřední deska je přístupná přes NKOD	Má úřad s rozšířenou působností
Bohumín	599051	00297569	Ano	Ano
Brandýs nad Labem-Stará Boleslav	538094	00240079	Ano	Ano
Česká Třebová	580031	00278653	Ano	Ano
Chotěboř	568759	00267538	Ano	Ano
Mladá Boleslav	535419	00238295	Ano	Ano
Moravská Třebová	578444	00277037	Ano	Ano
Nový Bor	561860	00260771	Ano	Ano
Podbořany	566616	00265365	Ano	Ano
Tábor	552046	00253014	Ano	Ano
Valašské Klobouky	585891	00284611	Ano	Ano
Boskovice	581372	00279978	Ne	Ano
Třeboň	547336	00247618	Ne	Ano

■ **Figure 2.5** Illustration of the List of municipalities page

2.3.3 List of municipalities

On this page, there is a list of municipalities. Each municipality in the list has its name, RUIAN, IČO, information on whether the associated municipality office published its official notice board on NKOD, and also information, specifying if the associated municipality office has extended competence. In the case that the municipality office has extended competence and does not publish its official notice board, then the column with the publishing information will turn red, specifying, that that municipality should publish its board. Additionally, the name of the municipality is also a link to a page with detailed information about it 2.3.4 and optionally its published board. Also, there is a search bar, that can be used to filter amongst the municipalities based on their name.

2.3.4 Detailed municipality view

Since on the *list of municipalities*, there is only some basic information, so that it is easy to orient in, there needs to be a page with more detailed information, for every one of the municipalities 2.6.

On the page, firstly there is some basic information about the municipality (same as on the *list of municipalities*). Then there is the published official notice board (only if it was actually published in the first place).

Each detailed view of a board has first some basic information about it, like its name, the name of its publisher, and a link to download it. It also has some basic graphs about the boards' data, if they are applicable to it. Then there is the list of all attached information. For every single piece of information, there are mandatory attributes like the name, link to a website, publish date, and date of relevance, but there are also some optional ones, like the reference number, description, and a list of attached documents.

In the case that a single municipality office published multiple boards, there is pagination to list through all of the boards. Based on my observations, this can happen for multiple reasons.

[O stránce](#) [Statistiky](#) [Obce](#) [Úřední desky](#) [Dokumenty na deskách](#)

Praha 12

RUIAN	547107
IČO	00231151

Úřední deska

Název	Úřední deska Praha 12
Název vydavatele dat	Městská část Praha 12
Odkaz ke stažení	Dostupný

Oznámení

Inzerát - technik	
Odkaz na web	Dostupný
Číslo jednací	Chybí
Datum vyvěšení	2022-05-04
Relevantní do	2022-05-20

■ **Figure 2.6** Illustration of the Detailed municipality part page for Prague 12

The two most probable ones are that either some publisher actually publishes one board twice (instead of updating the old one, they upload it as a new one and forget to delete the old one), or the boards are actually different. In the case of municipality offices, two different boards might be caused by the fact that one belongs to a municipality part office, which does not have its own IČO, so it uses the IČO of its superior office. In the case of other publishers, like state administration authorities or regional offices, it is because they also publish different kinds of boards.

2.3.5 List of official notice boards

The *data importer* can also import boards that were not published by any municipality office nor any municipality part office. These however cannot be accessible through the list of municipalities, because of that, there is also a page with a list of all official notice boards 2.7. Each row of the listing has the name of the board and the name and IČO of the publisher. Same as the *list of municipalities*, the user can filter records of boards, but in this case, it is by the name of the board, or the name of its publisher.

2.3.6 Detailed official notice board view

The detailed view of an official notice board 2.8 is very similar to the detailed view of a municipality 2.3.4. The only difference is that there is no information about the municipality at the top of the page. Additionally, since there is no information about a municipality, there cannot be any pagination between boards.

2.3.7 List of document

Like the list of municipalities and official notice boards, the list of documents 2.9 also has a search bar, this one however enables full-text search on the extracted text from documents.

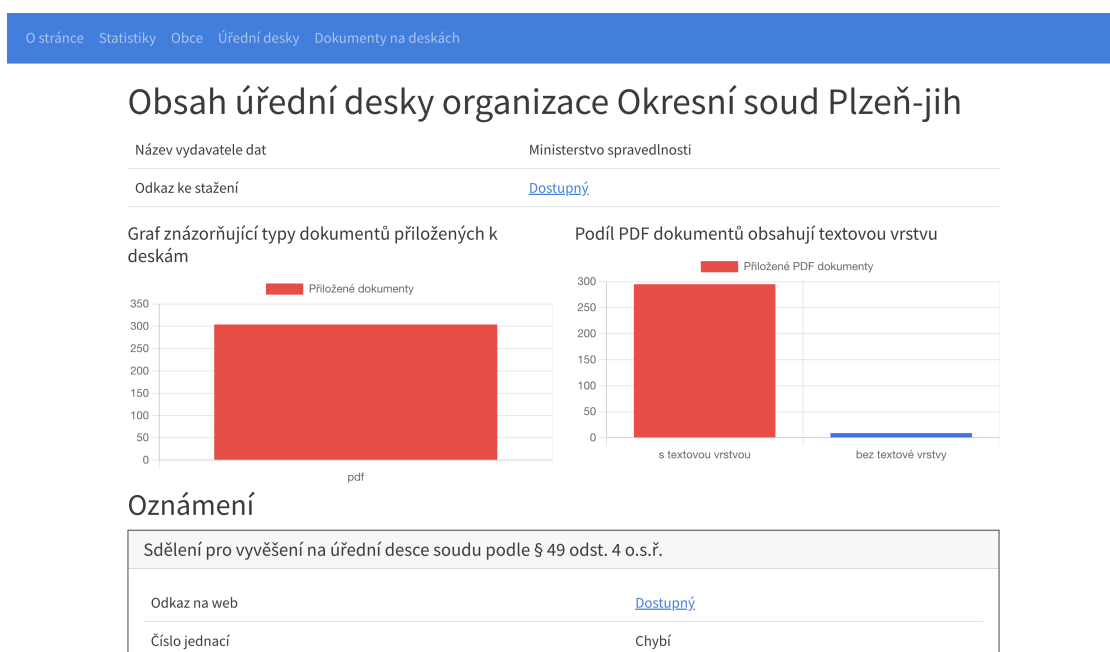
O stránce Statistiky Obce Úřední desky Dokumenty na deskách

Úřední desky

Název desky / vydavatele dat [Hledat](#)

Název	Název vydavatele dat	IČO vydavatele dat
Bezpečnostní odbor	Ministerstvo vnitra	00007064
Datová sada Úřední desky HZS	Ministerstvo vnitra	00007064
Datová sada Úřední desky MV	Ministerstvo vnitra	00007064
Datová sada Úřední desky PČR	Ministerstvo vnitra	00007064
Dočasné znovuzavedení ochrany vnitřních hranic	Ministerstvo vnitra	00007064
Dohody o změně veřejnoprávní smlouvy	Ministerstvo vnitra	00007064
Ekonomický odbor	Ministerstvo vnitra	00007064
Informace o úřednických zkouškách zabezpečovaných odborem personálním MV 1999	Ministerstvo vnitra	00007064
Kancelář ministra vnitra	Ministerstvo vnitra	00007064
Kancelář náměstka ministra vnitra pro řízení sekce ekonomiky a provozu	Ministerstvo vnitra	00007064
Kancelář náměstka ministra vnitra pro řízení sekce informačních a	Ministerstvo vnitra	00007064

■ **Figure 2.7** Illustration of the List of official notice boards page



■ **Figure 2.8** Illustration of the Detailed official notice board page

O stránce Statistiky Obce Úřední desky Dokumenty na deskách

Dokumenty

Hledat Hledat

Název	Odkaz ke stažení	Název vydavatele dat	Ukázka extrahovaného textu	Přípona souboru
(p) 16-C-183-2021-04-25-09-14-28-JS Sdělení pro vyvěšení na 9 odst.4 o.s.ř.-dokument.pdf	Dostupný	Ministerstvo spravedlnosti	Vyvěšeno dne: 25. dubna 2022 Sňato dne: ÚŘEDNÍ DESKA 16 C 183/2021 SDĚLENÍ pro vyvěšení...	pdf
(stejnopis)K-45-SI-223-2021-09-16-07-39-28-připisSI-vyhovněnížádosti+přílohy-dokument_redigováno.pdf	Dostupný	Ministerstvo spravedlnosti	OKRESNÍ SOUD V KARLOVÝCH VARECH Moskevská 17, 360 33 Karlovy Vary, tel.: 377867211, fax: 3...	pdf
(prvopis) 19-C-87-2022-04-14-13-56-54-vyvěšování na EÚD - sdělení4 o. s. ř. (DK)-dokument.pdf	Dostupný	Ministerstvo spravedlnosti	SDĚLENÍ (§ 49 odst. 4 o. s. ř.) Vyvěšeno dne: Sňato dne: 14. 4. 2022 16. 5. 2022 Adresát...	pdf
SIN 2_2015.pdf	Dostupný	Ministerstvo spravedlnosti	Okresní státní zastupitelství v Bruntále Partyzánská 11, 792 01 Bruntál VÁŠ DOPIS č.j.:	pdf
zveřejnění.pdf 4545643	Dostupný	Ministerstvo spravedlnosti	SIN 24/2015 Žadatelé – fyzické osobě – bylo poskytnuto anonymizované usnesení (v příloze). P...	pdf

■ **Figure 2.9** Illustration of the List of documents page

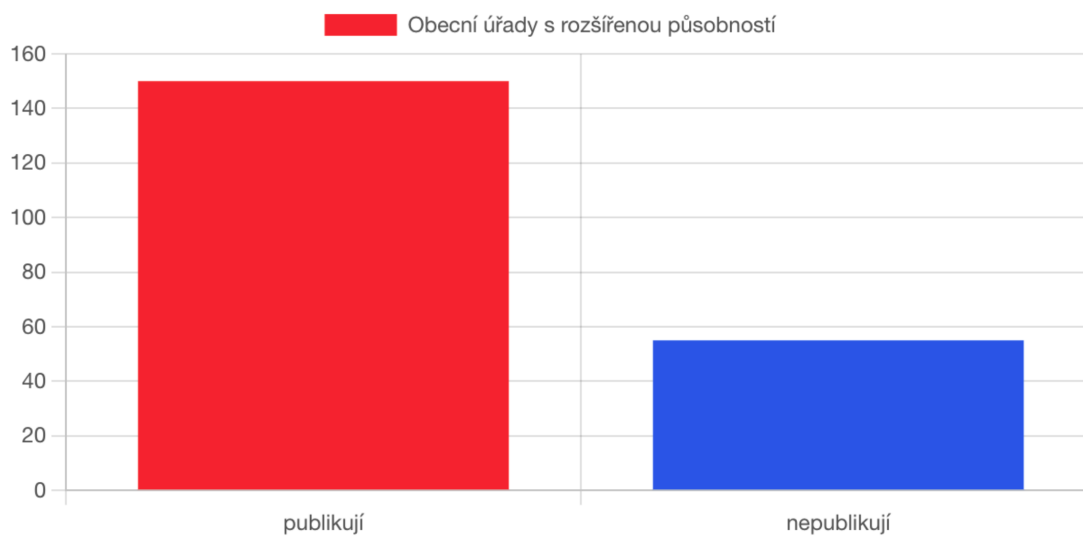
2.4 Statistics

On all pages, except for the *About page* and *Statistics*, there are listed some values of municipalities, official notice boards, information, and documents. Sometimes on these pages, some of these values can have a red or purple background. Red background means, that something is wrong, in that case, usually, the value of the given attribute is replaced by an explanation of what is wrong. For example, if there is some value missing that should be specified, based on the minimum specification from *OFNs*, like a name. Another common example is when there is a link to download a board or a document, but the link is either unreachable or it is not a download link, but it just leads to a website. As for the purple background, that one means, that something is not how it should be, but that doesn't necessarily mean it's wrong. This is used in most cases when the *data importer* has not yet processed the given record. So for example, it knows about some document, but it has not yet been downloaded and processed.

Used charts on the *Statistics* page 2.3.2, which is in its full size illustrated on the ??:

- How many municipality offices with extended competence publish boards – Municipality offices with extended competence have to publish their official notice boards and this chart shows how many actually do. 2.10
- Out of all boards published by municipality offices, how many were published by the ones with extended competence 2.11
- Number of errors by municipalities that violate the minimum specification 2.12
- Municipalities with the most amount of unreachable download URLs for documents 2.13
- Document file extensions 2.14
- How many of the PDFs contain a text layer and how many do not – When a PDF does not contain a text layer, in most cases it means, that it is a scanned document, without running any OCR run 2.15

Stav publikování obecními úřady s rozšířenou působností



■ **Figure 2.10** How many municipality offices with extended competence publish boards

- What municipality offices post the most PDFs without a text layer 2.16
Used charts on the **Municipality and Official notice board** pages: 2.3.2
- file extension document file extensions
- how many of the PDFs contain text layer (are not scanned)

Rozdělení desek obecních úřadu, podle typu úřadu

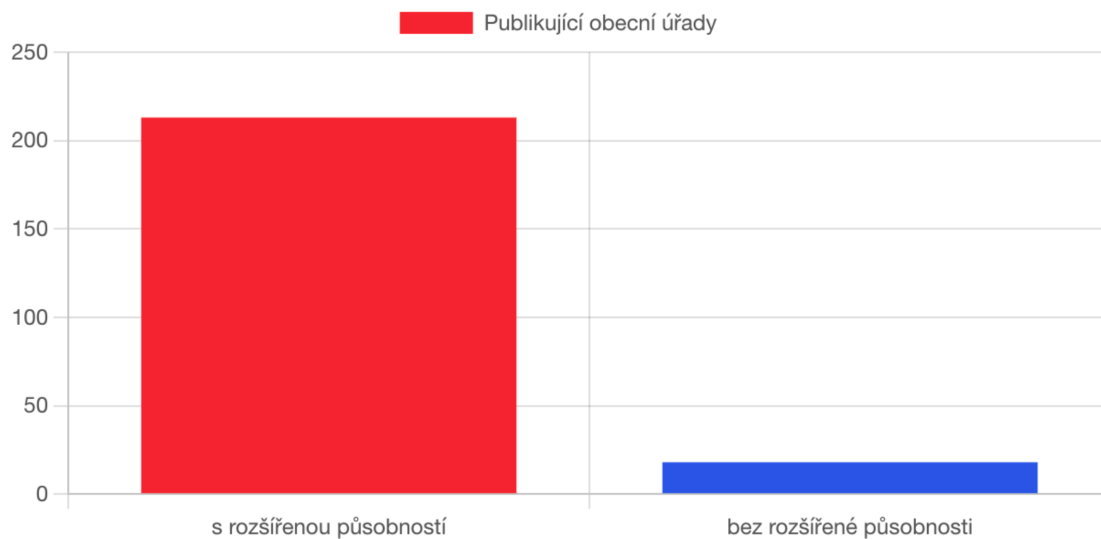


Figure 2.11 Out of all boards published by municipality offices, how many were published by the ones with extended competence

Desky obecních úřadů s nesplňují minimální povinnou specifikace

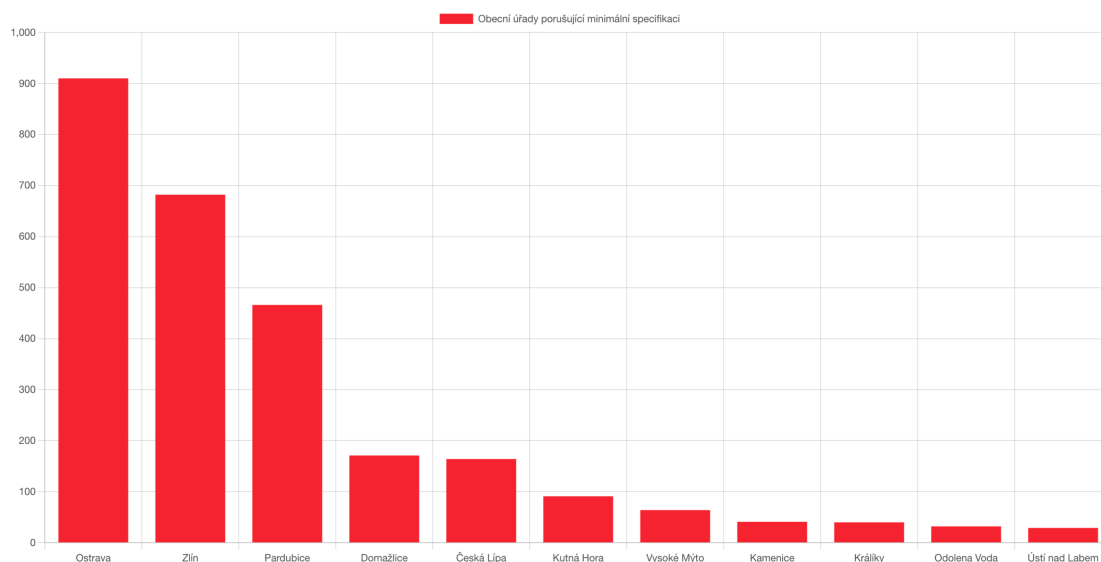
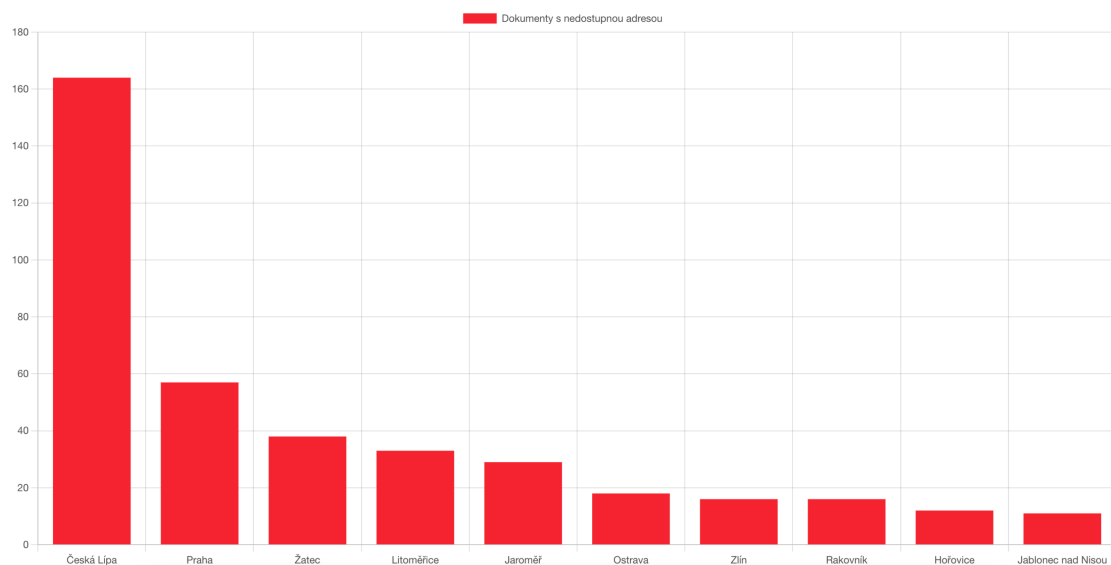


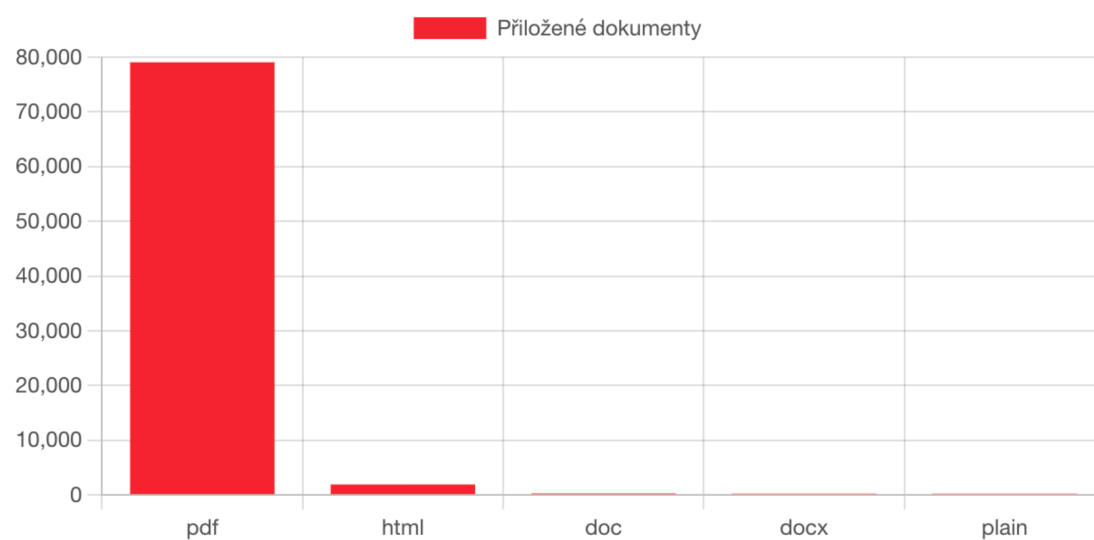
Figure 2.12 Number of errors by municipalities that violate the minimum specification

Desky obecních úřadů s nejvíce dokumenty s nedostupnou URL adresou



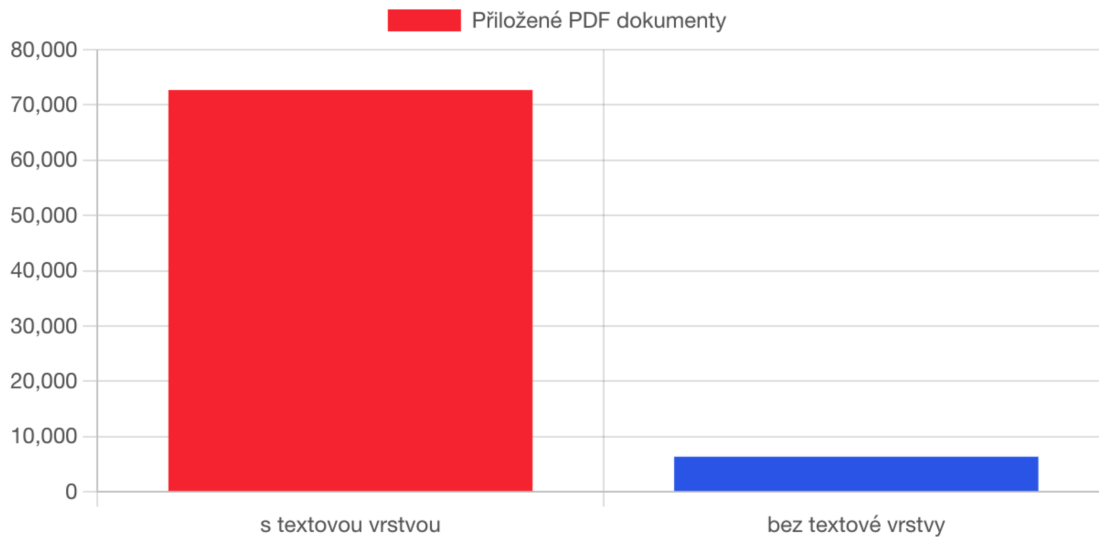
■ **Figure 2.13** Municipalities with the most amount of unreachable download URLs for documents

Typy dokumentů přiložených k úředním deskám obecních úřadů



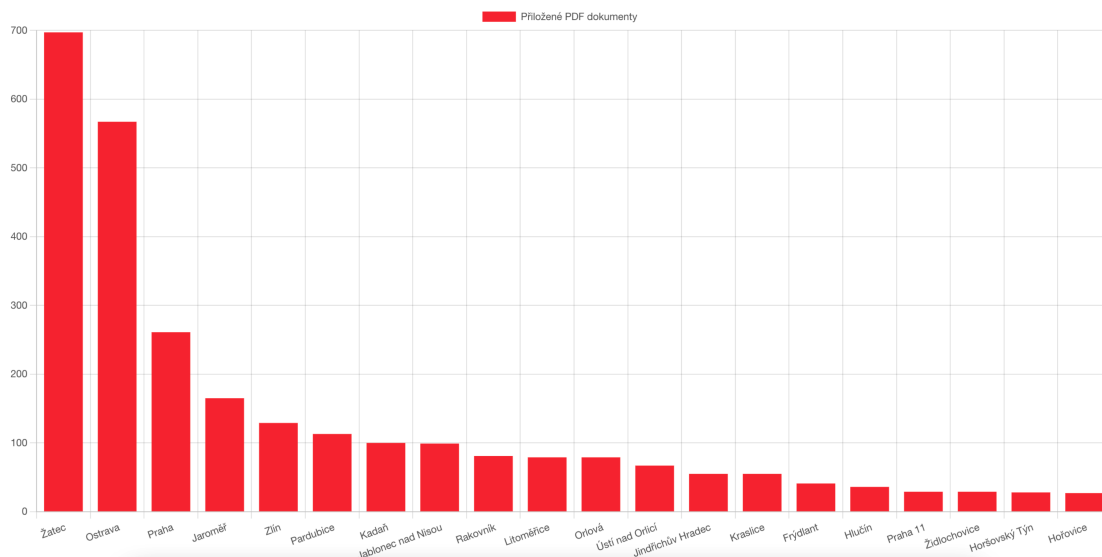
■ **Figure 2.14** Document file extensions

Podíl PDF dokumentů obsahujících textovou vrstvu



■ **Figure 2.15** How many of the PDFs contain a text layer and how many do not

Desky obecních úřadů s nejvíce PDF dokumenty, které nemají textovou vrstvu (naskenované dokumenty)



■ **Figure 2.16** What municipality offices post the most PDFs without a text layer

Implementation

This chapter focuses on the implementation of system, mainly the `data importer` and the `website`. First there is a description of some of the technologies, that are used.

3.1 Used technologies

- Python 3.10 – The `data importer` and the `website` back-end are implemented in Python. It was chosen out of personal preference. It is also the most popular language for data analysis and fast prototyping. As for the choice of version, 3.10 version was chosen, because it is currently the newest one and brings new features like pattern matching [15].
- PostgreSQL – It is used to store processed data. For doing full-text search, was used the `ts_vector`, which is part of PostgreSQL out of the box [14], which is one of the main reasons it was chosen.
- Flask – Flask is a micro-framework for building web application. [16]
- Flask-SQLAlchemy – Is a Python package, that allows object oriented mapping (ORM), which makes it easy to work with database records, directly in Python.
- Jinja – It is a templating engine, that has Python-like syntax and is used to generate HTML files [17].
- Docker – Docker is used to run parts of the system in separate containers [18].

3.2 System structure

In the design section 2.1, the system was split into 3 parts: `data importer`, `database` and `website`. However implementation wise, the `data importer` and `website` are more intertwined. Code-wise, they are part of the same Python package and use some of the same modules. They both use the same modules with database models.

The system is run in Docker containers using Docker-Compose. It is split into two containers. One is for the PostgreSQL database and the other is for the `data importer` and `website`.

3.3 Data importer

The stages of `import_all_data` are:

1. Create database tables
2. Get the list of municipalities and municipality parts
3. Map RUIAN to IČO
4. Mark municipality offices with extended competence
5. Get the list of available official notice boards metadata
6. Download boards + extract data
7. Download documents + extract text

The stages of `import_all_data` are:

1. Check if tables were created properly and there are some data
2. Get the list of available official notice boards metadata – different (slightly)
3. Download boards + extract data – different
4. Download documents + extract text – same

optional argument that specifies whether we want to import only boards published by municipalities, or all boards available

3.3.1 Create database tables

Database tables get created automatically by the Flask-SQLAlchemy Python package using `db.create_all()` command.

3.3.2 Get the list of municipalities and municipality parts

To get the list of municipalities and municipality parts, it first uses a SPARQL query on the SPARQL endpoint that is provided by NKOD. From this endpoint, it gets the download URL, for a JSON file with such a list. Then a REST GET request is used to download the file with the list of municipalities. You can see a shorter version in code block 3.1.

Using the SPARQL query to get the download URL, might appear to be a bit unnecessary since the URL is still the same `https://data.mpsv.cz/od/soubory/ciselniky/obce.json`. However it is not guaranteed, that the download URL will still be the same one in the future, so this way, it makes the `data importer` more "future-proof". The same goes for the download URL for a list with municipality parts.

3.3.3 Map RUIAN to IČO

The data from the list of municipalities, use RUIAN as an identifier of the municipalities and municipality parts. However, the imported data of official notice boards use IČO. This means that they currently cannot be mapped together.

A possible solution could be to compare the name of the municipality or municipality part, with the name of the publisher of the board. Unfortunately, this is not an ideal solution, since they are often in a slightly different format, for example for Prague 12, its municipality part name is *Praha 12*, but as a publisher of boards, it is *Městská část Praha 12*.

A better solution, that was eventually chosen is to use Wikidata, where municipalities have both of these records stored. To get the data from Wikidata I also use a SPARQL query.

```

{
  "polozky": [
    {
      "id": "Obec/554782",
      "kod": "554782",
      "nazev": {
        "cs": "Praha"
      },
      "okres": "Okres/3100"
    },
    {
      "id": "Obec/541575",
      "kod": "541575",
      "nazev": {
        "cs": "Žulová"
      },
      "okres": "Okres/3811"
    }
  ]
}

```

■ **Code listing 3.1** Example of a list of municipalities in JSON downloaded from NKOD.

```

SELECT DISTINCT ?municipality ?ruian
{
  ?municipality wdt:P31 wd:Q7819319 ;
                wdt:P7606 ?ruian
}

```

■ **Code listing 3.2** SPARQL query, that will return all municipality offices with extended competence

3.3.4 Mark municipality offices with extended competence

As written about in, out of all municipality offices, only those with extended power have to publish their official notice boards.

Just like the RUIAN to IČO mapping, to get this information, I also use a SPARQL query to get it from Wikidata. The query can be seen at 3.2 and an example of its response is at 3.3.

3.3.5 Get the list of available official notice boards meta-data

To download the official notice boards, I first need to get the download URLs for their files. To get them I again use a SPARQL query. The important part in the query is, the `confrontsTo` predicate. It has to be the link to the specification of boards on the OFN website, which for the current version as of writing is <https://ofn.gov.cz/%C3%BA%C5%99edn%C3%AD-desky/2021-07-20/>. A simplified version of such query is at 3.4

```

{"head":{"vars":["city","ruian"],"results":{"bindings":[
  {
    "city":{"type":"uri","value":"http://www.wikidata.org/entity/Q1789604"},
    "ruian":{"type":"literal","value":"597520"}
  },
  {
    "city":{"type":"uri","value":"http://www.wikidata.org/entity/Q1641781"},
    "ruian":{"type":"literal","value":"511021"}
  }
]}}

```

■ **Code listing 3.3** JSON response with a list of municipality offices with extended competence

```

PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX ofn: <https://ofn.gov.cz/>

SELECT ?datova_sada ?download_link WHERE {
  ?datova_sada dct:conformsTo ofn:úřední-desky\2021-07-20\ ;
              dcat:distribution ?distribution.
  ?distribution dcat:downloadURL ?download_link .
}

```

■ **Code listing 3.4** Example of a simple SPARQL query, that will return download URLs to all published official notice boards.

3.3.6 Download boards + extract data

Now I have the download URLs for the boards. To download them, a simple GET request will do the trick. An example of a downloaded board can be found at 1.1

Since the downloaded board should always be in the JSON file format, it can be converted into a Python Dictionary and from that, I can easily get some of the attribute values.

3.3.7 Download documents + extract text

As part of some of the downloaded official notice boards, there was a list of attached documents to information. And some of these documents have specified a URL, which can be used to download them. Just like with a download URL for the boards, I use GET requests to download these documents.

To enable full-text search on the text that is in the documents, it somehow needs to be extracted. The most common document file type is a PDF. Python package called `pdfminer.six` is used to extract text from the PDF documents. This package only extracts text from an actual text layer, meaning that when a PDF document is created by simply scanning a paper, it will not extract any text. This however can be used as an advantage, and when text extraction returns no text, it can be assumed, that it was scanned. The percentage of scanned documents is shown in the 2.15 chart.

The second most common document file format is HTML, which is actually a mistake. From observations of these occurrences, it's always because of one of two reasons. The first reason is, that the URL is not an actual download URL. The URL just leads to some regular page, where there might be some information about the document. The regular page has usually the real download URL. The second reason is a URL that is no longer valid, and it redirects to some generic page, saying invalid URL or access denied.

Text extraction is also supported on documents of type DOCX, for which is used the `docx2txt` Python package. For documents of types like PNG, JPG, and XLSX, it would make no sense to try and extract text from them, since they are usually just some extra documents, that go along with a PDF document, with no real text value. For example in the case of JPG and PNG, it is commonly a map of some land. And for XLSX it is usually just some table with budget calculations.

3.3.8 Modifications for `import_new_data`

3.3.8.1 Get the list of available official notice boards metadata

The only difference is that when the board already exists it will not be added again.

3.3.8.2 Download boards + extract data

Same as in the original version, the download will be called on all of the boards. However, if the board already existed, it will only add the new notices and deleted the ones, that are no longer on the board.

3.4 How to install and run

In the `impl` directory, there is a `README.MD` file, with all of the information on how to run the system. The steps are:

1. Make sure you have docker installed. Download link: <https://www.docker.com/products/docker-desktop/>

2. To start the website run `docker-compose up -d`
3. Open container shell `docker exec -it flask_app bash`
4. To import all data run `flask import_all_data` inside the container shell (This will take multiple hours)
5. To import new data since the last import, run `flask import_new_data` inside the container shell

3.5 Possible future improvements

Since things can be always better, there are of course some things that can be better in thesis as well. Some suggestions are:

- Improve text extraction from documents, by running it in parallel
- Run `import_new_data` automatically, preferably at least once a week.
- Calculate data for all of the charts at the end of `data importer` run, to improve website load times. Alternatively, the page views can be cached, for example with `Redis`.
- For text extraction from documents, support more file types.

Chapter 4

Summary

The goal of this thesis was to present and analyze a new open data about official notice boards. I was able to get a list of municipalities and links to their official notice boards. The next step was to analyze the data with regards to the Act No. 106/1999. The open data with some insights are presented on a user-friendly website that shows some characteristics of the data and allows search on them.

After I first got a list of all municipalities and municipality parts with their RUIANs, I got a list of links to all available official notice boards from NKOD, I found out, that from NKOD I was only able to get ICO of their publishers and not RUIAN. Because of that I also had to use Wikidata to get a list of mappings from ICO to RUAIN.

In the next part, I used the board links to download them. The downloading went smoothly, but then I had to process the data. During the data processing, I found out that many of them don't fully comply with the official format. For example, it is mandatory to specify board IRI, and URL is only optional however, most of these downloaded boards had only specified URL. Another problematic attributes were dates, where the problematic part wasn't specifying a date, but specifying that there is no date. Specifying that there is no date is still up to specification, but it has to be the correct way, which many boards did not have. Some boards also have some download links to attached documents. Some of these links do not even work in the first place, and when they do, some of the linked documents are not in a machine-readable format, like PDF which is made out of plain scanned pages.

After that, I used Flask and Jinja templates to create the website, which shows the boards, and their statistics and allows searching through the extracted documents. The default page shows some basic information about it, and it also has some graphs, that represent various statistics for all available boards. On the next page, there is a list of all municipalities and municipality parts. By clicking on any of those, you'll be redirected to a page, where you can see some info, statistics, and graphs of its board (if it has any). The last page lists all of the documents, and it supports a full-text search of their contents.

There are many possible improvements to this project. The most relevant would be to improve the data importing. Currently, when the `import_all` is run, it will download and process all currently available data all over again, however ideally this should be only done on the new data. Another good thing would be to increase test coverage. And not only with unit tests, but also with functional and integration tests.

The resulting project is published under the open-source GPL-3.0 license. It can be found on GitHub using this URL <https://github.com/opedatalabcz/official-notice-boards>

Bibliography

1. CZECH REPUBLIC. § 5a odst. 3 zákona č. 106/1999 Sb. o svobodném přístupu k informacím - znění od 01.02.2022 [online]. Zákony pro lidi.cz, 2010 – 2022 [visited on 2022-04-23]. Available from: <https://www.zakonyprolidi.cz/cs/1999-106#p5a-3>.
2. CZECH REPUBLIC. § 26 zákona č. 500/2004 Sb., správní řád - znění od 1. 1. 2021. [Online]. Zákony pro lidi.cz, 2010 – 2022 [visited on 2022-05-08]. Available from: <https://www.zakonyprolidi.cz/cs/2004-500#p26>.
3. NÁRODNÍ KATALOG OTEVŘENÝCH DAT. *Otevřená data a otevřené formální normy* [online]. 2020 [visited on 2022-05-08]. Available from: <https://data.gov.cz/%C4%8D1%C3%A1nky/otev%C5%99en%C3%A9-form%C3%A1ln%C3%AD-normy-01-%C3%BAvod>.
4. ŠKOP, Michal; KLÍMEK, Jakub. *Úřední desky* [online]. 2022 [visited on 2022-04-29]. Available from: <https://ofn.gov.cz/%C3%BA%C5%99edn%C3%AD-desky/2021-07-20/>.
5. EDESKY. *EDesky - O Projektu* [online]. 2014 – 2022. Available also from: <https://edesky.cz/o-projektu>.
6. NÁRODNÍ KATALOG OTEVŘENÝCH DAT. *Jednoduché zobrazení úředních desek pomocí Otevřené formální normy a Národního katalogu otevřených dat* [online]. 2022 [visited on 2022-04-23]. Available from: <https://ofn.gov.cz/%C3%BA%C5%99edn%C3%AD-desky/2021-07-20/aplikace/%C3%BA%C5%99edn%C3%AD-desky.html>.
7. NÁRODNÍ KATALOG OTEVŘENÝCH DAT. *Datová kvalita (nejen) v oblasti otevřených dat* [online]. 2022 [visited on 2022-04-23]. Available from: <https://data.gov.cz/datov%C3%A1-kvalita/>.
8. SOBOLEV, Mark. *Official city desk crawler* [online]. 2019 [visited on 2022-04-23]. Available from: https://github.com/opendatalabcz/notice-boards/blob/master/report/crawler_report_28.7.pdf.
9. CZECH REPUBLIC. § 4c odst. 1 zákona č. 106/1999 Sb. o svobodném přístupu k informacím - znění od 01.02.2022 [online]. Zákony pro lidi.cz, 2010 – 2022 [visited on 2022-04-23]. Available from: <https://www.zakonyprolidi.cz/cs/1999-106#p4c-1>.
10. CZECH STATISTICAL OFFICE. *Statistical metainformation system* [online]. 2022 [visited on 2022-04-23]. Available from: <https://apl.czso.cz/iSMS/en/home.jsp>.
11. WIKIDATA. *Wikidata:Introduction* [online]. 2013 – 2022 [visited on 2022-04-23]. Available from: <https://www.wikidata.org/wiki/Wikidata:Introduction>.
12. W3C. *SPARQL Query Language for RDF* [online]. 2008. Available also from: <https://www.w3.org/TR/rdf-sparql-query/>.

13. ELASTICSEARCH B.V. *Chapter 12. Full Text Search* [online]. 2022 [visited on 2022-05-11]. Available from: <https://www.elastic.co/guide/en/elasticsearch/reference/current/full-text-queries.html>.
14. THE POSTGRESQL GLOBAL DEVELOPMENT GROUP. *Chapter 12. Full Text Search* [online]. 1996 – 2022 [visited on 2022-05-11]. Available from: <https://www.postgresql.org/docs/current/textsearch.html>.
15. PYTHON SOFTWARE FOUNDATION. *Python 3.10.0* [online]. 2021 [visited on 2022-05-11]. Available from: <https://www.python.org/downloads/release/python-3100/>.
16. PALLETS. *Foreword* [online]. 2010 [visited on 2022-05-11]. Available from: <https://flask.palletsprojects.com/en/2.1.x/foreword/#what-does-micro-mean>.
17. PALLETS. *Introduction* [online]. 2007 [visited on 2022-05-11]. Available from: <https://jinja.palletsprojects.com/en/3.1.x/intro/>.
18. DOCKER INC. *Docker overview* [online]. 2013 – 2021 [visited on 2022-05-11]. Available from: <https://docs.docker.com/get-started/overview/>.

Contents of the enclosed media

impl	Source codes of the implementation part
└─ README.MD	Manual explaining how to run the application
└─ app	Source codes of the application
└─ tests	Tests of the application
└─ LICENCE	Licence of the project
text	Text of thesis
└─ thesis	Source codes of text of the thesis L ^A T _E X
└─ thesis.pdf	Text of thesis in a PDF file