

Interactive Online Reporting of Registry of Tuberculosis

Matěj Karolyi, Tereza Nováková, Petr Panoška,

Martin Komenda, Jiří Jarkovský

Institute of Health Information and Statistics of the
Czech Republic and Institute of Biostatistics and
Analyses, Faculty of Medicine, Masaryk University -
joint workplace,

Palackého nám. 4, 128 01, Praha 2, Czech Republic

Email: {karolyi, novakova, panoska, komenda,
jarkovsky}@iba.muni.cz

Jiří Wallenfels

National Tuberculosis Surveillance Unit, Hospital
Bulovka,

Budínova 2, 180 00 Praha 8-Bulovka

Email: jiri.wallenfels@bulovka.cz

Abstract—The existence of health registers at the national level allows for a systematic and continuous control of the state of healthcare, the digitalisation of certain services and the overall improvement of the state of care. The data collected in such registers are then analysed and visualised. Thanks to this, it is possible to effectively evaluate various indicators and thus pass on a comprehensive picture to all stakeholders involved in the health system - legislators, healthcare providers and patients. The Registry of Tuberculosis (RTBC) is one source of data. The article describes this data source as well as the methods that lead to the creation of a freely available online interactive tool. Published visualisations offer a view of the data from several angles and deal with various subtopics of tuberculosis. Finally, particular results and possible improvements are presented.

I. INTRODUCTION

THERE has been a recent trend to create complex visualization tools and data analyses from public healthcare datasets, which are stored inside various national registries. Especially for monitoring of crucial and serious diseases, the development of analytical and reporting tools over national health registries leads to an increase of knowledge and the research for problems of experts who are looking for current trends and who predict the possible trends. In the Czech Republic, such datasets are analysed and published by leading government institutions: Ministry of Health (MH), National Institute of Public Health (NIPH), Institute of Health Information and Statistics (IHIS), public health universities and others. Following information systems and have already been developed and published and can be considered as related work:

- **Czech Childhood Cancer Information System¹**— based on available data sources, Czech Childhood Cancer Information System improves the healthcare professionals' awareness of childhood cancers and promotes their education in this area. The target groups involve paediatricians, non-specialized paediatric departments of healthcare facilities and professionals in both basic and applied research, as well as civic associations of patients and their family members [1].

¹ <https://ccc-is.uzis.cz/index-en.php>

- **Comprehensive Cancer Care Network²**—a pilot model—established in two Czech regions: Vysočina Region and South Moravian Region. The pilot model covers all components of cancer care: from cancer prevention and organized screening programmes through standard diagnostic and treatment procedures to follow-up plans; specialized rare tumours-focused care as well as palliative are also included [2].
- **Map of Medical Equipment³**—the aim is to provide a clear visualization of information on the availability of individual categories of medical equipment at all health service providers that, according to the law, submitted the Annual Report on the instrumentation of a medical facility. The data used are based on the Program of Statistical Surveys of the Ministry of Health of the Czech Republic [3].
- **The National Database of Palliative Care⁴**— mapping and description of the end-of-life care in the Czech Republic. The portal aims to provide a complex view on the end-of-life issues and to contribute to the work of healthcare workers who make efforts to alleviate suffering and to improve the quality of life of tens of thousands of patients dying each year.
- **National Registration of Health Services Providers⁵**—the web portal provides a complete overview of all healthcare providers in the Czech Republic. It provides data on the profile and scope of care for medical facilities, contains information on the availability of medical facilities and other detailed information [4].

Tuberculosis disease (TB) and other mycobacterial infections have been monitored for a long time in the Czech Republic [5]. A newly developed visualization portal of the Registry of Tuberculosis (RTB) is part of this effort and belongs to the group of previously presented web portals. The portal is intended for experts in the field, but also for the

² <https://cccn.onconet.cz/>

³ <https://zt.uzis.cz/>

⁴ <https://www.paliativnidata.cz/index-en.php>

⁵ <https://nrpz.uzis.cz/>

general public, who can observe particularly the time trends of TB and other mycobacterial infections.

The web portals and online visualizations described above are based on data from the Czech National Health Information System (NHIS). NHIS comprises dozens of health registries and information systems and its existence is given by Czech legislation. Processing and publishing of health data for the purpose of obtaining clear information about the extent and quality of provided health services belong to the main services provided by NHIS. From the perspective of online data presentation describing selected domain of NHIS, the proper interpretation of given information must be in compliance with two aspects: (i) data from health registries must be processed, aggregated and visualized anonymously (with no direct identification of any patient record), (ii) potential combination of descriptive attributes in a form of dynamic filters over available datasets like sex, age, and location must never lead to show too low total number of cases per geographic unit (district or town).

Nowadays, interactivity is one of the most required features of web-based applications guaranteed and produced by MH and IHIS (e.g., connected with epidemiological data [6]), which present various results and data overviews [1]–[3]. Modern layout, attractive user interface, overview customization, and data filtering bring an unique way to properly deliver given information to the final user/target group. With the respect to this approach, an interactive online reporting of the tuberculosis registry has been designed, developed and deployed.

A. Czech Registry of Tuberculosis

The Registry of Tuberculosis (RTB) is a part of Czech public health registries maintained by MH, IHIS, NPIH and Regional Public Health Authorities (RPHA)⁶. In the Czech Republic, reporting of all detected cases of tuberculosis to the RTB is obligatory by law. Therefore, the registry includes all individuals with active tuberculosis or other mycobacteriosis and all persons treated in groups of active and inactive tuberculosis or other mycobacteriosis which were diagnosed in the Czech Republic. As the Information System of Bacillary Tuberculosis (ISBT) has recently become an integral part of this registry, RTB is a very complex data source for this type of infection. Data from RTB are processed by IHIS and a statistical summary publication reflecting the current epidemiological situation in the Czech Republic is published once per year, entitled Basic Overview of Tuberculosis Epidemiology in the Czech Republic. Together with the National Tuberculosis Surveillance Unit, IHIS also maintains contacts with international organizations.

II. METHODS

The whole process of design and development is divided into different steps, which are iteratively repeated if needed. During the whole period we keep in mind the verified methodology for data mining called Cross Industry Standard Process for Data Mining (CRISP-DM) [7] in a domain-

specific environment [8]. This helps us to standardize our endeavour and optimize the delivery of the final product. Our development process often leads to the detection of low data quality. Therefore, we involve techniques for the improvement of the underlying dataset. This ensures continuous fixing of data entries even in the source Registry of Tuberculosis.

A. Data description

Data entering the web application are transmitted from data analysts to software developers in csv format. There are also several supporting files created, including the list of variables and its character (binary, numeric, string, etc.) and possible values, a logical framework displaying the allowed and forbidden usage of filters in given analyses, and a summary of required graphic outputs for each analysis.

The final dataset itself includes numerous descriptive variables, such as demographic data (age, gender, region of residence, etc.), important report dates (for example year of reporting or year of death) and a set of selected clinical parameters (type of infection, diagnosis by ICD-10 code, etc.). Apart from the main dataset, two separate datasets exist as a data source for the analyses of treatment results and susceptibility to antituberculotic drugs. Only filtering by year of reporting is available for these analyses.

The combination of those datasets is quite unique worldwide (according to our knowledge only few countries present their TB data in online and publicly accessible form). Therefore, our newly developed online reporting tool could have ambition to become an inspiration to similar teams in other countries which want to present their own TB data.

B. Methods of analysis

To display real numbers of cases of TB and other mycobacterioses, it is necessary to apply standardised pre-filtration processes. Based on specified diagnosis (by ICD-10 coding system), year of reporting and several other variables were identified as key ones to describe relevant (correct and complete) records in RTB data. Two standard statistical methods were subsequently applied, based on the character of individual analysis.

Firstly, the frequency analysis is used to determine the absolute numbers of cases, which are displayed for instance in analyses “Time trends—cases” or “Time trends—deaths from TB”. In analyses where percentages were preferable, the relative frequency analysis is employed, such as “Age distribution”, “Country of birth” or “Mycobacterial examination”, where the percentages provide a better perspective.

C. Technological background and deployment

From a technical point of view, the portal consists of a relational database, backend (server-side) framework and frontend (client-side) framework. A complete architecture is shown in Fig. 1.

⁶ <https://www.uzis.cz/res/f/008271/tbc2018-en-a2b.pdf>

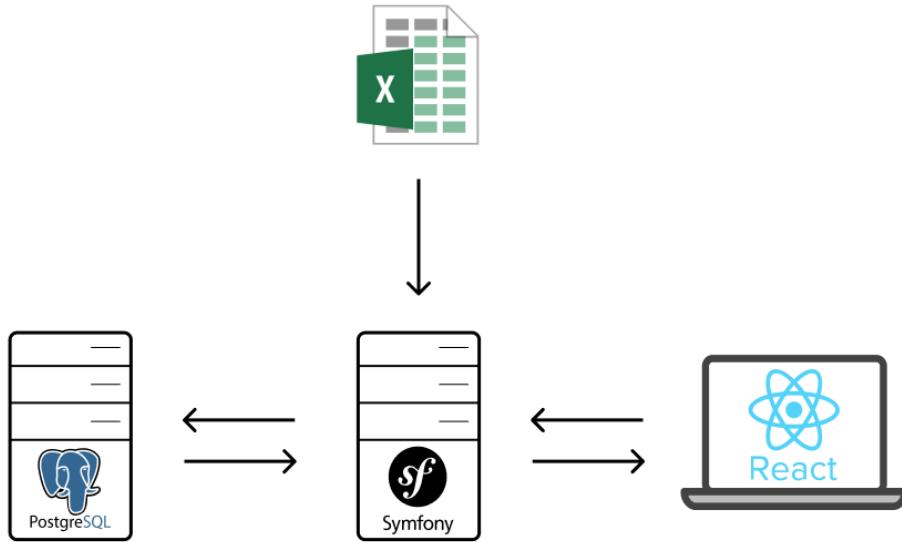


Fig. 1 Portal architecture schema

PostgreSQL is used due to our previous experience and simultaneous usage in other web services. Selection of the backend platform was also based on existing knowledge and current technical environment setup; thus we use the Symfony framework. The choice of frontend framework is driven by the need of dynamic content rendering and usage of several javascript libraries for data visualization. For this reason, React was selected. Fig. 1 also demonstrates that the data discussed earlier are served to Symfony, where their import to the database is performed. Data is mapped to PHP entities with the Doctrine Object Relational Mapper (ORM) framework, which is a suitable solution for transmitting data between database and Symfony.

When the data is ready to be queried by the backend server, API endpoints are implemented for each analysis, in which data preprocessing is done. Filters selected by the user are applied and data aggregation and normalization are performed during this process. Final data are served in JSON format to the client.

Key aspect of choosing React and its major advantage is re-rendering content according to user analysis type and filters' selection (application state) without reloading the whole page. Typically, when the user loads an analysis page (view), visualization and also static data are served to the client and the page is rendered. For the next user interactions with the view, only the visualization data are responded by the server and visualization components are updated. Most of the components are charts from the NVD3 javascript library, which has a high-level API for convenient components updates. As a result, users can work with the analysis without unnecessary reloading, and chart components smoothly update their values, which is an important user experience improvement.

III. RESULTS

The Interactive Registry of Tuberculosis Data Viewer is a web-based tool that allows health professionals to gain an overview of the state of the disease from various perspectives. This tool is freely accessible at <https://tbc.uzis.cz/> via a web browser. All currently available analyzes are available without the need to log in and at the time of publication of the article contain data from 2000 to 2019.

Current structure of the portal allows further expansion with potential further analyzes and graphical outputs. The currently available ones are further supplemented by static outputs directly on the IHIS.

A. Portal structure

The portal is divided into several sections, where users can find the content connected with the topic of tuberculosis disease and other mycobacterial infections in the Czech Republic.

Information about the registry

Basic static pages contain introductory information and an overview of the context in which the portal has been developed. Part of this overview is also a basic introduction to the Registry of Tuberculosis, whose establishment and operation are regulated by a governmental decree.

One of the important outputs are yearbooks of tuberculosis. The publication "Basic Overview of Tuberculosis Epidemiology in the Czech Republic" has been published regularly by IHIS CR since 2016 (with data for 2015) and indirectly follows the publication "Tuberculosis and Respiratory Diseases" which contains data from 1960 to 2014. It is a selection of the most important overview tables characterizing the occurrence of reported cases of

tuberculosis and other mycobacterioses in the Czech Republic. These documents are available on the IHIS CR website⁷.

Data viewer

Each analysis has its own separate visual presentation. Two analyses are available only for the Czech Republic as a whole, but most of them offer a region-based view on data. Analysis outputs can also be narrowed through several filters, which comprise basic demographic variables as well as selected clinical parameters.

Current analyses (views) over the following topics:

- Time trends—cases
- Time trends—deaths from TB
- Age distribution
- Nation of birth
- ICD diagnoses
- Proportion of previously treated patients
- Tests for mycobacterial infection
- Treatment results
- Susceptibility to antituberculotics

B. Selected data visualizations

In this paper, only a subset of the portal content is shown and described; three the most dissimilar were selected for demonstration. The rest is available and freely accessible online. Visualization outputs are generated and displayed online; however, the user is also allowed to download figures and tables to their computer.

Analýza: Časové trendy - případy



Fig. 2 Trend of TB disease from 2000 to 2018 in a group of patients older than 60 years (Czech only)

⁷<https://www.uzis.cz/index.php?pg=registry-sber-dat--ochrana-verejneho-zdravi--registr-tuberkulozy#publikace>

Time trends—cases

First analysis called “Time trends—cases”, which shows the number of detected cases of TB and other mycobacterial infections, is displayed as a time series line graph. That allows the user to explore the development of these infections over the whole period 2000–2019. Even more illustrative is the map, showing abundance of the diseases in Czech regions.

The output of this analysis shows a time series of cases reported in the Registry of Tuberculosis. Using the tabs, it is possible to switch between graphical and tabular overviews for the Czech Republic and for individual regions, and the cartogram. The results can be further filtered using selected parameters from the "filtering" panel. Fig. 2 shows the line chart of the TB incidence trend from 2000 to 2018 in a group of patients over 60 years.

Age distribution

When looking at the output of the “Age distribution” analysis, represented by a histogram displaying the relative numbers in 5-year age categories, one can easily determine the age groups in which TB occurs with the highest frequency. When we divide the whole country to particular regions, a set of 14 boxplots is displayed (one boxplot for each region).

The same as in the previous case, it is possible to switch between graphical and tabular overviews for the Czech Republic and for individual regions. The results can be further filtered on demand. Fig. 3 shows a histogram of all age groups of Czech TB patients in a reduced time period.

Analýza: Věkové rozložení

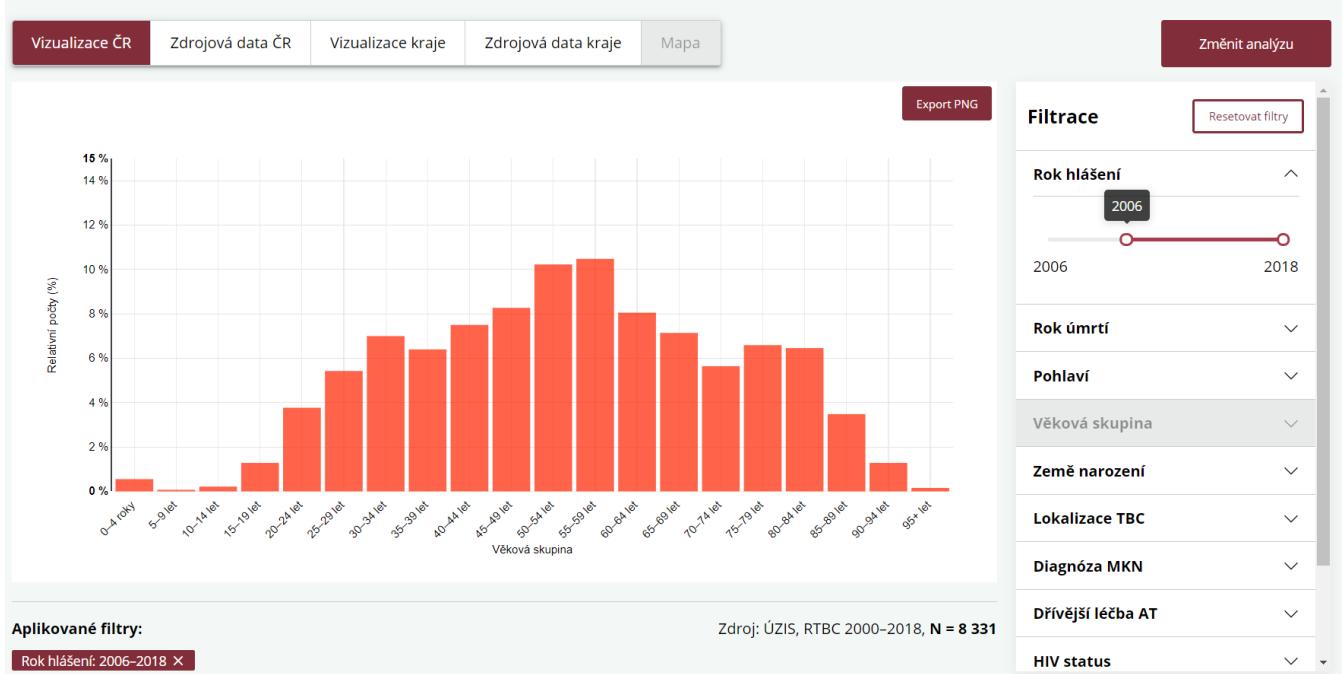


Fig. 3 Histogram of all age groups of czech TB patients (Czech only)

Tests for mycobacterial infection

The analysis “Tests for mycobacterial infection” is available for TB cases only and shows a barchart of relative frequencies of possible examination results—culturally positive TB, culturally positive TB verified from sputum or laryngeal swab (LS), and sputum smear microscopy-positive TB. The output of this analysis allows the user to compare the total number

of cases of tuberculosis in the Registry of Tuberculosis with the number of cases of tuberculosis according to possible results of mycobacterial tests.

Fig. 4 shows the analysis of regions according to results of mycobacterial testing in 2018. The presented figure does not contain all regions.

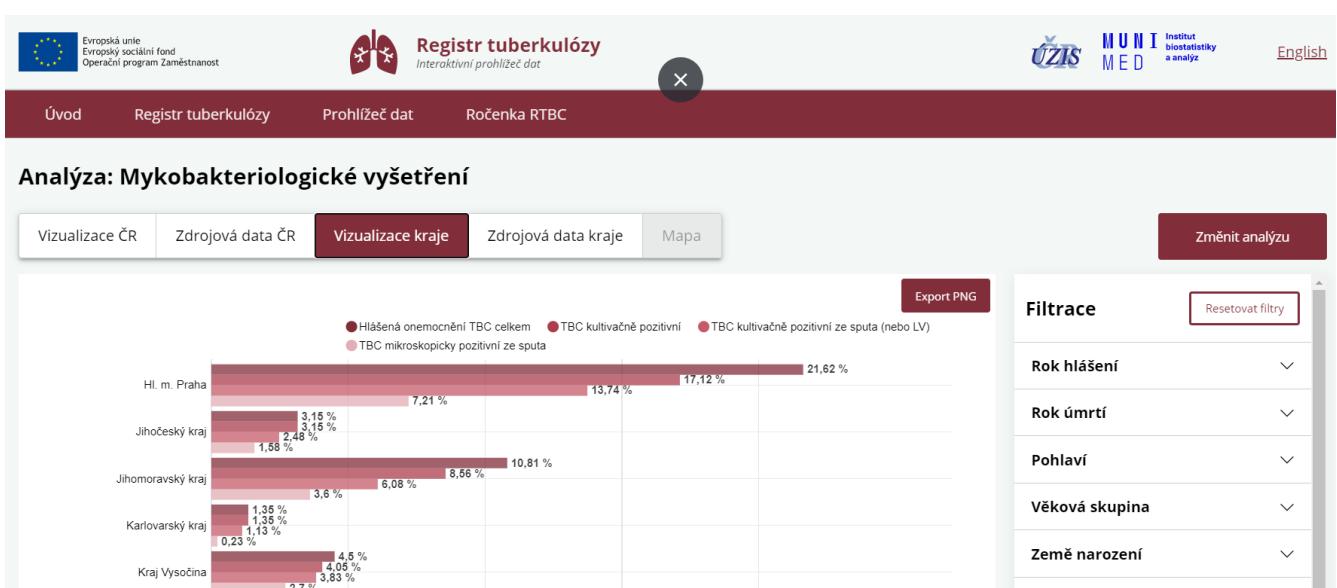


Fig. 4 Analysis of regions according to mycobacterial testing results (Czech only)

II. DISCUSSION

In this article, we describe a web application tool for the exploration of data from the RTB. Users of this portal can choose between various analyses—from the basic ones based on demographic characteristics to some more advanced ones, which are suitable for experts, such as doctors or other experts. Worldwide, only few countries present their tuberculosis data and to our knowledge no country reports the cases of other mycobacterial infections in a form of web portal. This online tool is therefore unique in combining the two infections together in one application.

However, there are several limitations as well. Firstly, it is possible to access data from the year 2000 onwards only. When we compare this tool with similar applications based on TB data from the Netherlands⁸, one can see a longer time period to display. Secondly, the outputs of analyses are currently presented in absolute numbers and/or percentages only. We plan to add the possibility to present data as relative incidence. Thirdly, in comparison to the Dutch portal, data about latent TB infection (LTBI) are missing—this is due to the fact that the surveillance system of LTBI has still not been established in the Czech Republic. Also, the offer of possible outputs seem to be more extensive on the Dutch website, but since the Czech one has been intended to address the general public too, we decided to show the most important parameters only.

Another way to improve the comparability of data between regions or even other countries can be achieved if some the values are given as relative (e.g., per 100 000 inhabitants).

Currently, the web portal is intended to be updated once per year by adding the definitive data for the previous calendar year. Alternatively, we consider more frequent updates with the use of preliminary data or even “rolling” updates reflecting the continuous reporting to RTB in real time.

We believe that in future this web portal could be considered as an example of data presentation for other infectious diseases that are mandatorily reported to the Information System of Infectious Diseases⁹.

III. CONCLUSION

As a result of multidisciplinary cooperation, which included three important groups—experts in the field of tuberculosis

and other mycobacterioses, data analysts, and software developers—we developed a freely available tool displaying visualizations of data from the National Tuberculosis Registry in the Czech Republic, which has a great potential to be beneficial for both professional and general public. The tool is based on modern technologies, which makes future long-term sustainability as well as scalability and extensibility much easier.

The content of the portal has already been reviewed several times by the professional community; their feedback was taken into account and the important points and comments were also implemented retrospectively. At the time of writing this paper, the content is only available in the local language version. We are planning a translation to English.

The data update is scheduled for an annual cycle. Within the datasets, there is no intention to change the data structure over the years; only new annual datasets will be added. However, if necessary, it will be possible to adapt to possible changes in the source registry (e.g., as a result of adjustments to data collection from healthcare providers).

REFERENCES

- [1] D. Krejčí *et al.*, ‘Webový portál pro analýzu a vizualizaci dětských onkologických dat dostupných v ČR’, 2019, Accessed: Jul. 01, 2020. [Online]. Available: <https://www.med.muni.cz/veda-a-vyzkum/publikacni-cinnost/1576436>.
- [2] L. Dušek, J. Mužík, M. Karolyi, M. Šalko, D. Malůšková, and M. Komenda, ‘A Pilot Interactive Data Viewer for Cancer Screening’, in *Environmental Software Systems. Computer Science for Environmental Protection*, Cham, 2017, pp. 173–183, doi: 10.1007/978-3-319-89935-0_15.
- [3] M. Karolyi, J. Krejčí, J. Ščavnický, R. Vyškovský, and M. Komenda, *Tools for development of interactive web-based maps: application in healthcare*. Václav Skala - UNION Agency, 2019.
- [4] ‘Legislation - ÚZIS ČR’. <https://www.uzis.cz/index-en.php?pg=nhis--legislation> (accessed Jul. 01, 2020).
- [5] *Tuberkulóza a respirační nemoci* 1998. Ústav zdravotnických informací a statistiky ČR, 1999.
- [6] M. Komenda *et al.*, ‘Complex Reporting of the COVID-19 Epidemic in the Czech Republic: Use of an Interactive Web-Based App in Practice’, *J. Med. Internet Res.*, vol. 22, no. 5, p. e19367, 2020, doi: 10.2196/19367.
- [7] F. Martínez-Plumed *et al.*, ‘CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories’, *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2019, doi: 10.1109/TKDE.2019.2962680.
- [8] O. Niaksu, ‘CRISP data mining methodology extension for medical domain’, *Balt. J. Mod. Comput.*, vol. 3, no. 2, p. 92, 2015.

⁸ <https://www.tbc-online.nl/eng/index.html>

⁹ <https://www.uzis.cz/index.php?pg=registry-sber-dat--ochrana-verejneho-zdravi--informacni-system-infekcni-nemoci> (czech only)