Charles University in Prague

Faculty of Mathematics and Physics

BACHELOR THESIS

Tomáš Beňák

Triplification and presentation of statistical crime data according to Linked Data principles

Department of Software Engineering

Supervisor of the bachelor thesis:	RNDr. Jakub Klímek, Ph.D.
Study programme:	Informatika
Specialization:	IOI

Prague 2014

I would like to thank Jakub Klímek, the supervisor of this thesis, for his supreme support and guidance.

I would like to thank Bc. Vladimír Stolín, the police officer who provided me with the source data about crime in the Czech republic.

Finally I would like to thank my brother Ondrej Beňák who provided me with a lot of information about the laws of the Czech republic, especially the Criminal law.

I declare that I carried out this bachelor thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In .... date .... signature of the author

Název práce: Triplifikace a prezentace statistických dat o trestné činnosti podle principů Linked Data

Autor: Tomáš Beňák

Katedra: Katedra softwarového inženýrství

Vedoucí bakalářské práce: RNDr. Jakub Klímek, Ph.D., Katedra softwarového inženýrství

Abstrakt:

Klíčová slova: Linked Data, trestná činnost

Title: Triplification and presentation of statistical crime data according to Linked Data principles

Author: Tomáš Beňák

Department: Department of Software Engineering

Supervisor: RNDr. Jakub Klímek, Ph.D., Department of Software Engineering

Abstract: Linked data as a set of data publication principles and technologies has become increasingly popular. The number of datasets is growing and the Web of Data is becoming larger and more interconnected. Despite the growing trend, there are many domains still not covered by the Linked Data Cloud. In the Czech republic, the amount of published Linked Data is even smaller. There are several publishers, linked.opendata.cz being one of the important ones, providing a rather small set of interconnected datasets. As their dataset portfolio grows, more opportunities to connect with those datasets arise. One interesting domain still not covered is crime statistics. This thesis describes the whole process of searching for and discovering of potential data sources, retrieving the data and publishing them as Linked Data in RDF format along with a descriptive vocabulary. A separate part of the thesis deals with the development of the demo-application that would demonstrate the usability of the published Linked Data.

Keywords: Linked Data, crime

Contents

Introduction

Cílem práce je získat a v otevřeném a strojově čitelném formátu RDF podle principů Linked Data publikovat data o trestné činnosti na území České republiky. Publikovaná data budou veřejně dostupná a jejich užitečnost je demonstrována v ukázkové aplikaci. V úvodu práce je popsána motivace pro publikaci dat v kontextu existujících datových sad. První kapitola popisuje sběr a přípravu dat z různých zdrojů a použité technologie. Druhá kapitola vymezuje cílovou skupinu uživatelů a stanovuje požadavky na výsledné řešení. Třetí kapitola popisuje zvolené řešení (použité technologie, alternativní řešení). Čtvrtá kapitola obsahuje zhodnocení navrženého řešení a diskusi jeho alternativ. Pátá kapitola popisuje ukázkovou aplikaci z uživatelského hlediska. Přílohy obsahují programátorskou dokumentaci ukázkové aplikace a všechny diagramy a tabulky.

1. Introduction

1.1 Linked data as a way of publishing

In recent years, Linked Data has become increasingly popular as a data publishing paradigm. Its simple set of principles, use of well-defined and understood standards, ability to represent the data universally and a growing suit of underlying technologies has made it a popular means of publishing structured data on the Web.

L inked data is nowadays used to publish the data from many topical domains including commerce, book publishing, science or the democratic process.

Fond Otakara Motejla In the Czech republic, a non-governmental organisation Fond Otakara Motejla tries to act as a platform for the data publishers and data consumers, such as application developers. It acts as a bridge reaching to state institutions in their effort to make those institutions publish the results of their work in an open format, so that it can be consumed. Open data is a data publishing paradigm, making the data publishers provide a publicly available format for the published data and provide the data on regular basis. Linked Data as a publishing paradigm implements the Open Data paradigm and is used to publish Open Data in the Czech republic, for example by linked.opendata.cz. Some of the most notable results of the activities of the Fond Otakara Motejla include the publishing of the dataset of checks performed in restaurants and some more. (See the web page).

M onitoring the democratic process and the process of the administration of the state finances are the two fields in particular interest even of common people. In the UK, there is the TheyWorkForYou initiative that monitors the activities of the representatives in the British Parliament.

As a part of the LOD initiative the Public Contracts datasets and ontology have been developed in order to enable the public to review the issued public contracts as well as to provide a mechanism to pair the offer and demand in this field, effectively bypassing the potentially corrupted administration.

In the Czech republic, there have been similar initiatives. The data publisher linked.opendata.cz puvlished the dataset about the work of the Parliament along with some other datasets.

linked.opendata.cz Linked.opendata.cz published quite a few Linked Data datasets with their value increasing over the time as more and more datasets are being created and interlinked. Some of the notable datasets published by linked.opendata.cz are the Municipalities dataset and the Laws dataset. One of the domains still not covered is crime statistics in the Czech republic.

2. Requirements and Goals

The Crime statistics dataset would become a new member of the *linked.opendata.cz* datasets family. It is intended to be served on a *linked.opendata.cz* server and thus residing in the *linked.opendata.cz* URL namespace.

2.1 Requirements

Several requirements have been proposed:

on detail and completeness The published dataset is intended to be the most detailed and complete source of public crime statistics data there is. Of course some data cannot be made available for public, typically personal data.

ability to stay up-to-date There should be an easy-to-use mechanism to perform the dataset update when new data becomes available.

public The dataset should be made publicly available through standard Linked Data technologies. This means it should be discoverable and its data should be made accessible for querying.

linking to other datasets The dataset would be connected to other datasets where such links emerge naturally. These connections enable the Crime dataset to become a part of the global data-space, the Web of Data.

extensibility The Crime Ontology (vocabulary) developed to describe the Crime dataset structure should be easily extensible when the source data structure changes.

usefulness The dataset should prove itself useful.

2.2 Goals

Building upon the requirements we set the goals of this thesis:

the data source Gather the most complete and detailed crime statistics data. Exclude the personal data.

- Publish the gathered data as an RDF dataset in an RDF store and make its data available at a SPARQL endpoint.
- Develop an ontology or a vocabulary to describe and maintain the structure and semantics of the published data. The ontology will be built reusing the existing well-known vocabularies to make it easier to comprehend, use and extend.

update mechanism An update mechanism will be developed to make it easy for the dataset maintainer to perform the dataset update in case new data is available.

- The dataset will be linked with resources in other datasets, for example DBPedia and linked.opendata.cz datasets.

usefulness An Android demo-application will be created to demostrate the usability and usefulness of the dataset. The application will provide crime statistics based on the user's current location.

2.3 Potential Users

There are three groups of potential users of both the dataset and the demo-application.

application developers The primary target audience of the Crime dataset. The developers can issue interesting queries on the data, mashing them up with related data using the geography location as the key.

taset and ontology creators The other datasets creators may want to reuse a part of the Crime vocabulary when talking about crime.

Android users Every user of a device powered by a supported version of Android is a potential user of the demo-application. The main value for such user would probably be entertainment.

3. The Crime Statistics Dataset

TODO Notes about different versions of the Czech Criminal Law.

3.1 Source forms

Field number	**Field name**	**Description**	**Value**
*01	identification number	Identifies the filled FTČ form.	structured field with several subfields described in a helper table 3.3
*02	stage of crime	Classifies the stage reached when committing a crime.	Codelist 3.1
*03	type of offense	Classifies the crime according to its relation to extremism.	Codelist document 3.1.1
*04	tactical-statistical classification	Classifies the crime using various internal criteria of the ESSK.	Codelist document 3.1.1
*05	crime committed department	The code of the base department on the territory of which the crime has been committed.	Codelist document 3.1.1
*06a	committed on the street	Determines if the crime had been committed on the street.	yes / no (1 / 2)
*06b	monitored site	Records if a crime had been committed on the monitored site.	yes / no (1 / 0)
*06c	crime scene location type	Classifies the type of location where the crime had been committed.	Codelist 3.2
*07a	weapon use	Records information about whether a crime offender had used a weapon and how and classifies the consequences if any.	Codelist document **??**

3.1.1 Codelists

Stage of Crime

This codelist classifies a crime according to in which phase it has been detected.

Value	Description
1	preparation
2	attempt
3	completed

Table 3.1: Stage of crime

Crime Category

This codelist document defines a classification of crimes and a classification of crime offenders.

Extremist crimes classification The crime classification classifies a crime according to its relation to extremism. It is used in FTČ field *1 and in FZP field 28 for each individual crime of an offender.

Several extremist crime categories are recognized, based on the target of crime like crimes against religion groups, members of a nation or race. There are some other ”non-targetting”, more general crime categories, like terrorism or spectator violence. [TODO grammar]

There is an additional crime categorization applied atop of the described classification. It puts crimes into categories based on their severity. According to the new Czech Criminal Law there are less serious crimes (”přečiny”) and more serious crimes (”zločiny”). In addition to this distinction there is a third group of crimes. It applies to crimes that have been committed prior to 31. 12. 2009 and as such have been qualified using the old version of the Czech Criminal Law and are neither ”přečiny” nor ”zločiny”.

Crime offenders classification The second classification defined by this codelist document is the classification of crime offenders. It is not used by the current version of the FZP form, but have formerly been used in the field 12 of the FZP form.

It classifies a crime perpetrator according to his or her relation to extremism. If there is such a relation, several extremism categories are distinguished like right or left wing political extremism or religious extremism.

The location of the original HTML document is provided in the References section.

Tactical-statistical crimes classification

This codelist document defines the so-called tactical-statistical crimes classification (TSK).

Using various internal criteria considering different aspects of a crime such as its legal qualification, the crime target and other circumstances, a crime is assigned a TSK classification value. Each TSK classification value is explicitly linked to qualifying laws using a list of references to the sections of the Czech Criminal Law.

The TSK classification is hierarchical. Each category represents the generalization of the contained TSK classification values or whole other crime categories. There is for example a crime category group called Property crimes containing various crime categories dealing with various kinds of theft such as Thefts - general or Thefts - burglary. Thefts-burglary crime category then contains a direct listing of individual TSK classification values associated with differents subtypes of burglary.

There are two versions of this document. The first version is used to classify the crimes qualified using the old version of the Czech Criminal Law. The second version classifies the crimes qualified by the current version of the Czech Criminal Law.

The location of the original HTML document is provided in the References section.

Police departments classification

This document contains the descriptions of individual departments of the Police of the Czech republic and their hierarchy.

For every police department there is a

- code
- name
- flag whether the department has a territory
- flag whether the department's territory is under surveillence
- description of a change that occured on this department (new department, department cancelled, department moved, department's name has changed)
- a contextual note on the department's change, if any (previous name, the department that takes over the cancelled department's agenda etc.)

Organization of the departments of the Police of the Czech republic is hierarchical. The structure of the hierarchy is build upon the existing hierarchy of the territorial and administrative units in the Czech republic.

At the top level there are fourteen (previously eight) regional headquarters and various departments that operate outside the boundaries of the regional headquarters.

Regional headquarters operate on the territory of a whole region. An example would be KŘP STŘEDOČESKÉHO KRAJE, the regional headquarters of the Středočeský region. Some of the departments that belong to regional headquarters are organizationally directly under the headquarters. These departments don't have the territory assigned. An example of such department is ETŘ KŘPS.

However most of the departments are further organized into smaller organizational units, police districts. Each police district has its own headquarters and the territory of a police district is usually the one of the corresponding district of the Czech republic. An example of a police district is ÚO Benešov corresponding to the Benešov district.

The territory of a police district is divided between the local departments of the district. Usually such a local department operates on the territory of one or more municipalities or town districts. Conversely there can be several local police departments for a single municipality. This means that generally if a crime is committed on the territory of a local department, one cannot link the crime to a municipality because that information is not provided.

Some local departments don't have the territory assigned. Only a department with the territory can be used to fill in the field *5 of the FTČ source form to provide a code of the department on the territory of which a crime has been committed. An example of a police district local department with the territory would be OOP Benešov. An example of a police district local department without the territory would be OOK ÚO SKPV BENEŠOV.

There is another group of departments that operate outside of the context of the regional headquarters. Those departments are generally the departments with special agenda like fighting organized crime, police inspection, human trafficking and other. These departments are organized into organizational units according to the type of agenda they perform, not according to the territorial nor administrative division. An example of such specialized police department would be ÚOOZ ODB. TERORISMU A EXTR., which specializes in fighting terrorism and extremism. This department is located in the PČR ÚOOZ SKPV organizational unit.

Codes of departments The codes of the police departments are structured in a way that follows the hierarchy of the departments. The regional headquarters has a four digit code. The first two digits are the logical code itself, other two digits are zeros intended as padding to a four character length. KŘP STŘEDOČESKÉHO KRAJE has the code 0100.

The code of a police district consists of four digits as well. The first two digits correspond to the two-digit prefix of the code of the parent regional headquarters. The other two digits are the code of the police district within the region. ÚO Benešov has the code 0101.

A police department (local, directly under regional headquarters, central etc.)

has a six-digit code. The first four digits correspond to the code of the parent organizational unit such as the district or the region. These four digits are then followed by the two-digit local code of the department within the parent organizational unit. ETŘ KŘPS has the code 010000. OOP Benešov has the code 010110. ÚOOZ ODB. TERORISMU A EXTR. has the code 200407 and the parent PČR ÚOOZ SKPV organizational unit has the code 2004.

Changes of departments Number of changes can occur during the time that affect the codelist of police departments. Generally once a department is assigned a code, no other department can be assigned the same code in the future. A department's name can change. In this case we are provided with both the new and the previous name. An existing department ceases to exist. Usually another department takes over the cancelled department's agenda. We are provided with the codes of both departments in question. A department migrates across police districts or even regions. It is assigned a new code within its new parent district. A new department may be introduced.

There have been some major organizational changes in the past being the result of changes in administrative and territorial division of the Czech republic. Previously there have been eight regions, each with its own regional headquarters. Now there is a total of fourteen regions, each with its own regional headquarters. The underlying police districts remained the same, they only moved across the newly established regional headquarters.

The location of the original HTML document is provided in the References section.

Crime location

This codelist classifies the location where the crime has either been committed or reported. There are several location categories recognized by the ESSK listed in the table bellow.

Value	Description
1	railways
2	highway
3	subway
4	remote area
5	park
6	populated area
7	cottage colony
8	abroad
9	urban settlement
A	Internet
B	other computer network
0	other

Table 3.2: Crime location

Weapon use

This document aims at providing the classification of crimes

3.1.2 Helper tables

Subfield name	Description	Value
kraj	The region code.	two characters (digits)
okres	The county code.	two characters (digits)
útvar	The code of a base department in the context of its parent county department.	two characters (digits)
Č.J.	The reference number.	9-character string composed of digits
rok	The year when the FTČ form has been filled.	two-digit year representation
poř.č.	The ordinal number of the FTČ form.	two digits

Table 3.3: Identification number - subfields

Codelist value	Description

Table 3.4:

4. Ontology

In order to represent the source data in a structured manner intended for further processing we had to design a data model. We chose to abstract the individual crimes and their perpetrators as observations of a statistical dataset. Therefore we would use the standard W3C Data Cube vocabulary. There were several reasons why representing the data in a statistical dataset was suitable. TODO: reasons

- reason a
- reason b

We identified the total of three distinct logical datasets in the source data. One of them we would use to hold the crime records, one would be dedicated to hold the perpetrators records and the third one would be a RDF linkset represented as a Data Cube dataset.

Conclusion

Bibliography

List of Tables

List of Abbreviations

Attachments