



Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

Publication and usage of official Czech pension statistics Linked Open Data

Jakub Klímek^{a,*}, Jan Kučera^b, Martin Nečaský^a, Dušan Chlapek^b

^a Department of Software Engineering, Faculty of Mathematics and Physics, Charles University, Malostranské náměstí 25, 118 00 Praha 1, Czech Republic

^b Department of Information Technologies, Faculty of Informatics and Statistics, University of Economics, Prague nám. W. Churchilla 4, 130 67 Praha 3, Czech Republic

ARTICLE INFO

Article history:

Received 30 January 2017

Received in revised form 23 June 2017

Accepted 17 September 2017

Available online xxxx

Keywords:

Pension

Statistics

Data Cube Vocabulary

Linked Data

ABSTRACT

Linked Open Data (LOD) principles are known for a decade now and there are thousands of LOD datasets of variable quality and importance. They often come from academic research projects, which show how LOD can be published, how it can be useful, etc. However, these projects last only for a few years and when they end, the datasets often cease to be maintained. What is needed is to convince the owner of the original data, e.g. an organization in public administration, to keep on publishing LOD on their own even when the project ends. Therefore, it is noteworthy and admirable, when this happens. In this paper we describe how the Czech Social Security Administration (CSSA) publishes official pension statistics as LOD as part of their day-to-day operation, which was jump-started by an applied research project. The data is modeled using the Simple Knowledge Organization System (SKOS) vocabulary and the RDF Data Cube Vocabulary (DCV). It is published as RDF data dumps, as a SPARQL endpoint and using IRI dereferencing for semantic web technologies power users. For journalists and other users without knowledge of these technologies, the data is also published as CSV files and visualizations generated from the LOD. We show how the data is reused in applications and how it contributes to statistical indicators in combination with other LOD.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

During the last 5 years, open data has become a hot topic in the Czech Republic as more and more organizations in the public administration and municipalities start to publish their data. As founders of the OpenData.cz initiative¹ we helped some of them with the process of opening their data from scratch. Most of the organizations publish their data only as CSV or XML files, i.e. as 3-star data according to the 5-star Open Data Deployment Scheme.² Few of them publish their data as 5-star data, i.e. as Linked Open Data (LOD). One of the first organizations which published its data as LOD was the Czech Social Security Administration (CSSA). Among other datasets, the CSSA publishes its official Czech pension statistics as LOD.

1.1. The Czech social security administration

The Czech Social Security Administration is an organization which is part of the Czech public administration. It collects

social security premiums, which include pension insurance, sickness insurance and contributions to the state employment policy. In addition, the CSSA covers social security issues for over 8.5 million clients. In total, it manages 750 billion CZK each year.

Initially, the CSSA published only statistical yearbooks as print-ready PDF files and paper booklets.³ These yearbooks, however, contained valuable information about various types of pensions, numbers of pensions paid, etc., with a great potential for publication as open data. Also included was information regarding demography published by the Czech Statistical Office.⁴

In 2014, we started working on an applied research project, which had a goal of experimental verification of how the data from the CSSA statistical yearbooks could be published as Linked Open Data and combined with other datasets to gain interesting new insights. This research project was executed between January 2014 and June 2015. It was managed and delivered by a joint team of the University of Economics, Prague (UEP) and an IT company that acted as an industrial partner in the project. The CSSA was not directly involved in the project, however it provided consultations

* Corresponding author.

E-mail address: klimek@ksi.mff.cuni.cz (J. Klímek).

URL: <http://www.ksi.mff.cuni.cz/~klimek> (J. Klímek).

¹ <https://opendata.cz>.

² <https://5stardata.info/en/>.

³ In Czech: <http://www.cssz.cz/cz/o-cssz/informace/informacni-materialy/statisticke-rocenky.htm>.

⁴ <https://www.czso.cz/csu/czso/home>.

Table 1Examples of items from the *pension kinds* code list.

Code	Label
SD	Old-age pension in accordance with Section 30 of Act No. 155/1995 Coll
ST	Old-age pension in accordance with Section 31 of Act No. 155/1995 Coll
IT	Disability pension of the 3rd grade in accordance with Sections 38 and Section 39(2)(c) of Act No. 155/1995 Coll. or relevant provisions of previous regulations, including disability pension IM in accordance with Section 42 of Act No. 155/1995 Coll. or relevant provisions of previous regulations
VM	Widower's pension in accordance with Section 48(2) of Act No. 155/1995 Coll. or relevant provisions of previous regulations

to the project team that were essential for understanding of the pension statistics data and the related code lists.

In 2015, the CSSA had a decision to make. They needed to decide whether or not they will publish data from their statistical yearbooks at least in the form of the source, Excel spreadsheets, in addition to the PDF and paper versions. At that time we suggested that they could go even further and publish the data as CSV files to achieve the level of 3-stars according to the 5-star deployment scheme of open data. However, their IT department wondered why not go all the way up to the level of 5-stars, i.e. to the level of Linked Open Data. Therefore, a decision was made and a plan was drafted to publish the yearbook datasets primarily as LOD, and also in CSV, which was to be generated from the LOD representation for those who do not know SPARQL, RDF and the used vocabularies. Activities leading to publishing of the statistical yearbook datasets as LOD started in May 2015. These activities included setting up the technical environment described in this paper as well as defining accountability and responsibilities of the relevant organizational units of the CSSA. Representatives of the UEP participated in these activities as invited experts. The CSSA published the first datasets at the beginning of November 2015. Since then, the CSSA has published 63 datasets⁵ as LOD.

1.2. Contributions

The contributions of this paper include the description of the LOD preparation and publishing environment used in the CSSA and its evaluation, survey of the resulting datasets, description of how the datasets are modeled using SKOS and the RDF Data Cube Vocabulary and to which datasets they are linked. Finally, we show the impact of the published datasets and how they can be easily combined with other LOD datasets to gain new, interesting insights into the data.

1.3. Outline

The rest of this paper is structured as follows. In Section 2 we describe the input data as it was gathered internally in the CSSA. In Section 3 we describe the vocabularies used to model the output LOD. Section 4 contains description of the data linkage. In Section 5 we describe the technical environment used for the data publication process. In Section 6 we overview the CSSA datasets published so far. In Section 7 we evaluate the publishing environment according to various criteria. In Section 8 we demonstrate the impact of the published datasets by presenting some of the known usages of the published data. The section demonstrates the usability and usefulness of the published data in practice. In Section 9 we survey related work and finally in Section 10 we conclude.

Table 2Examples of items from the *statistical reasons for expiration* code list.

Code	Label
01Z	Pensioner died
04Z	Expired disability consequently expires entitlement to IP, ID, IT or V, VM
08Z	Pension expires because different pension is awarded
12Z	Pension granted erroneously

2. Input data overview

The data to be published is not available in any nice machine-readable format. It is basically a set of PDF files combining texts and tables formatted for printing, which makes them useless for further machine processing. They need to be converted to a better machine-readable representation which, however, makes the publishing process much harder. Fortunately, the CSSA provided the source, print-ready Excel files, from which the tables in the PDF files were created. This inconvenience is caused by the definition of the yearbook publishing process inside the CSSA. If an error is found in the materials ready for printing, it is corrected there, and not propagated into the primary data sources. Therefore, the print ready Excel files and the final PDF files are the only authoritative sources of data which can be published. There is no additional metadata documents available. This is an inefficiency and source of inconsistencies in the data handling processes inside the CSSA. However, until this is resolved, the work has to be done based on the print ready files. There are two main kinds of input statistical data: code lists, and the data itself.

2.1. Code lists

There are three types of code lists used in the statistical yearbook datasets. Code lists of the first type are specified explicitly in the PDF files of the yearbook. Implicit code list values are present in the data, but there is no definition. Finally, there is a code list for regions in the Czech Republic, which is external to the CSSA. Some of the code lists are quite extensive, so only samples are shown here.

2.1.1. Explicit code lists

First, there are four code lists explicitly specified in Chapter 3 of each of the yearbooks in PDF, both in Czech and in English. Their RDF version is created manually.

The *pension kinds* code list specifies pension kinds defined by the Czech legislation. Examples of code list items can be seen in Table 1.

The *statistical reasons for expiration* code list contains various reasons for why a pension is no longer paid out. Examples of code list items can be seen in Table 2.

The *statistical reason for grade change* code list states various reasons for change of a pension. The code list items can be seen in Table 3.

⁵ In Czech: <https://data.cssz.cz/web/otevrena-data/katalog-otevrenych-dat>.

Download English Version:

<https://daneshyari.com/en/article/6950446>

Download Persian Version:

<https://daneshyari.com/article/6950446>

[Daneshyari.com](https://daneshyari.com)