LINKED OPEN DRUG DATA FROM THE HEALTH INSURANCE FUND OF MACEDONIA

Milos Jovanovik, Bojan Najdenov, Dimitar Trajanov

Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University

Skopje, Republic of Macedonia

ABSTRACT

Information dissemination has always been in the focus of the computer science research community. New ways of information and data representation, storage, querying and visualization are being constantly developed and upgraded. Linked Open Data represents a concept which offers a comprehensive solution for information and data dissemination. It accomplishes this by aiming towards two things: to represent data in an open, machine-readable format, and to interlink data from heterogeneous repositories in a way which allows a large variety of usage scenarios for both humans and machines. On the other hand, health also represents a domain of high interest in our research community. In order to provide use-case scenarios for publishing and using healthcare data in Macedonia, we generated a dataset of five-star Linked Open Data, based on the data provided and published by the Health Insurance Fund (HIF) of the Republic of Macedonia. In this paper, we describe the process of transforming the data available at the HIF website, into data published in an open format, and interlinked with data from the DrugBank domain.

I. INTRODUCTION

The basic idea behind the Open Data concept is that data which can be considered public should be available in a raw and machine-readable format, for the purposes of use, reuse, republishing and redistributing, with little or no restrictions. When public datasets are published in an open format, they can be used for building useful applications which leverage their value and offer different use-cases for the interested parties [1]. Furthermore, these datasets can contribute to the overall development of the society, by both boosting the ICT business sector with new business value and providing the stakeholders with new functionalities [2].

On the other hand, the concept of Linked Data provides mechanisms for interlinking data from different repositories distributed on the Web, in order to provide better data usage and querying scenarios [3]. Linked Data, in a way, represents a synonym to the Semantic Web, since its main goal is interlinking data from the Web by their meaning. The Linked Data techniques rely on identifying resources with URIs, providing data about these resources and connecting them to other resources on the Web, by using standards such as the Resource Description Framework (RDF) [4].

The Linked Open Data Cloud [5] consists of interlinked datasets which have been published in Linked Data format, across the web. The datasets contain data from various domains: media, geographic data, publications, user-generated content, government, cross-domain, and health and life sciences.

Health is a research area which, when it comes to data and information, is one of the main topics of interest for computer scientists. Over the years, many different approaches for representation, storage, querying and visualizing of health data have been developed. The new techniques employed by the Linked Open Data community offer new ways for covering these areas of interest for health data. This is one of the main reasons we decided to work with data from the Health Insurance Fund of the Republic of Macedonia.

II. RELATED WORK

Numerous efforts have been made worldwide so far for transforming healthcare data into Linked Data. The most notable are the Linking Open Drug Data (LODD) project, LinkedCT, Open Biological and Biomedical Ontologies (OBO), and the Semantic Web Health Care and Life Sciences Interest Group at W3C.

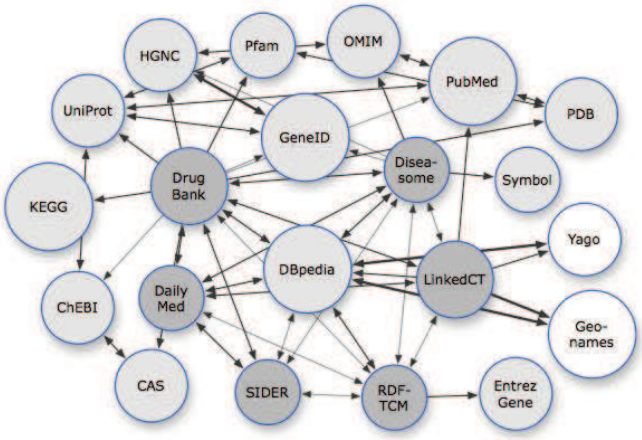

Figure 1. Part of the LODD Cloud.

The LODD project[1] is focused on interlinking data about drugs already existing on the Web [6]. The data ranges from impact of the drugs on gene expression, through to results of clinical trials. The aim of the project is to enable answering of interesting scientific and business questions by interlinking previously separated data about drugs and healthcare. As part of their work, they have collected datasets with over 8 million RDF triples, interlinked with more than 370.000 RDF links (Fig. 1).

One of the datasets, which is a part of the LODD cloud, is DrugBank[2]. It provides RDF data about drugs, such as chemical, pharmacological and pharmaceutical information, taken from an existing base[3] of drug data on the web. The DrugBank RDF dataset contains over 766.000 RDF triples for 4.800 drugs. Because of its size, we decided to use this dataset as a reference point for the drug data from the Health Insurance Fund which we describe, publish and interlink.

LinkedCT[4] is a project aimed at publishing clinical trials data in a Linked Data format [7]. They transform existing clinical trials data into RDF, discover semantic links between the records themselves, and link to other data sources such as PubMed[5], as well. The datasets from LinkedCT are also part of the LODD project (Fig. 1).

OBO Foundry[6] – the Open Biological and Biomedical Ontologies project, is a collaborative effort involving biology researchers and ontology developers who work together to develop a set of design principles for ontology development in the biomedical domain.

The Semantic Web Health Care and Life Sciences[7] is an interest group at the World Wide Web Consortium (W3C), comprised of experts from around 30 W3C member organizations: research centers, universities, companies, health institutions, etc. Its mission is to develop and support the use of the technologies of the Semantic Web in the fields of healthcare, life sciences, clinical research and translational medicine [8].

The Open Data and Linked Data research activities in Macedonia so far include the development of a Crime Map for the Republic of Macedonia [9], as well as the opening and linking of data from the Universities in Macedonia [10]. Apart from these, there have not been Open Data and Linked Data activities involving healthcare data from Macedonia.

III. Linked Open Data from the Health Insurance Fund

The Health Insurance Fund of the Republic of Macedonia is an institution which is responsible for regulating and managing the public services for primary healthcare, specialist healthcare, and hospital healthcare. Additionally, the Fund along with other government institutions regulates the list of drugs which are covered by the health insurance, and defines the referent (nominal) prices for certain drugs.

With this position in the society, we believe that the data which the Fund works with is of high importance, and there would be a great benefit of opening their public data in RDF, and interlinking it with other datasets from the LOD and LODD clouds.

A. Public Data from the Fund

The Health Insurance Fund of the Republic of Macedonia has been publishing their public data on a regular basis on their website[8]. These data contain information about healthcare services and their prices, statistics about the rate of usage of hospital beds, reports from the inspections in the public and private healthcare institutions, financial data about the Fund, insurance information, referent drug prices, private and public healthcare institutions which the Fund works with,

[1] http://www.w3.org/wiki/HCLSIG/LODD
[2] http://wifo5-03.informatik.uni-mannheim.de/drugbank/
[3] http://drugbank.ca
[4] http://linkedct.org/
[5] http://www.ncbi.nlm.nih.gov/pubmed
[6] http://obofoundry.org/
[7] http://www.w3.org/blog/hcls/
[8] http://www.fzo.org.mk/

etc. The Fund has not yet published its data on the official Macedonian Open Government Data portal[9].

Although the data from the Fund's website can be technically considered as Open Data, they are mainly published in PDF and Excel formats, making them only one-star and two-star data, according to the Linked Open Data rating system [4]. In order to leverage the usability of the public data from the Fund, we decided to transform them into five-star Linked Open Data: to first transform them into RDF, and then interlink them with data from other publicly available datasets from the LOD and LODD clouds.

B. *Ontology*

The Fund has published their public drug data in various datasets, which contain different sets of information. These datasets hold information about the brand name, the generic name, the manufacturer, the referent price, the packaging, the strength, and the dosage form for drugs. Additionally, each drug is identified by an ID generated by the Fund, as well as a globally identifiable ATC code, used for classification of drugs and controlled by the World Health Organization.

In order to transform and represent the drug data in RDF, we needed an ontology. Following the best practices for ontology development, we decided to re-use already existing drug ontologies. In the process of choosing an ontology for re-use, we had to bear in mind the interlinking part of the process, which meant that we need an ontology used by a drug dataset which we would connect our data to, later in the process. With this in consideration, we decided to use the DrugBank RDF repository and its ontology. The DrugBank ontology contains the class 'drugs', which represents the drug entities. It also contains relations for the ATC code, the generic name and the brand name. We used the 'drugs' class along with the 'atcCode', 'genericName' and 'brandName' properties (Fig. 2).

However, we still needed properties for describing the other drug information, not covered by the DrugBank ontology. Therefore, we developed our own ontology: the HIFM ontology (Fig. 2). The HIFM ontology contains its own class for drug type entities, 'Drug', seven datatype properties and 'similarTo' as an object property (Fig. 2).

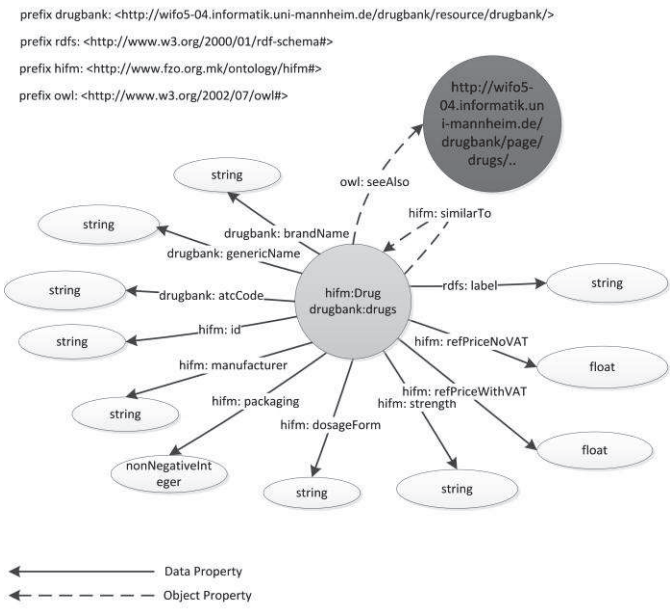

Figure 2. The HIFM Ontology.

Along with the properties taken from DrugBank, and the properties defined in our HIFM ontology, we use the 'rdfs:label' and 'owl:seeAlso' properties. The 'rdfs:label' property is used to point to the generic name of the drug, whereas 'owl:seeAlso' is used to link the drugs from our HIFM graph with drugs from DrugBank. This will be elaborated in more details further in the paper.

C. *Mapping the Data from CSV to RDF*

The next step was to map and transform the public data from CSV to RDF. For this, we decided to use the Virtuoso Universal Server[10], which provides mechanisms for data transformation and management, for various types of data, including the Semantic Web standard representation format, RDF. It serves as a Linked Data server, as well, and allows local and remote data querying with the Semantic Web query language, SPARQL.

The mapping process consisted of two steps. First, we imported the CSV files (generated from the Excel files available on the website of the Health Insurance Fund) into relational databases in Virtuoso. Then, with the use of R2RML[11], the mapping language for transforming RDB data into RDF data, which is also a part of Virtuoso, we created RDF Views over the relational databases. These RDF Views allow data management with the use of the technologies of the Semantic Web, such as querying with SPARQL, over data

[9] http://opendata.gov.mk/

[10] http://virtuoso.openlinksw.com/

[11] http://www.w3.org/TR/r2rml/

which resides in standard relational databases. In this step we used our HIFM ontology, as well as parts of the DrugBank ontology which were previously discussed. As an identifier of the drugs we chose the ID value, assigned to the drugs by the Fund. Each of the drugs was set to be both of 'drugbank:drugs' and of 'hifm:drug' RDF type, and the values for the ATC code, the generic and brand name, the dosage form, the strength, etc., were described using the DrugBank and HIFM properties (Fig. 2).

D. Transforming the RDF Data into Linked Open Data

Once we had all of the drug data into an RDF graph in Virtuoso, we proceeded with interlinking the drugs among themselves and with other drugs available in the LOD and LODD clouds.

For the purpose of interlinking the drugs from the Fund between themselves, we created a property in the HIFM ontology, called 'similarTo' (Fig. 2). This property has the purpose to link Drug A to Drug B (and vice-versa), if their first seven characters from the ATC code match. Even though the ATC codes should have seven digits, the ATC codes which the Fund assigns to the drugs in Macedonia contain ten digits. These additional three digits are used for marking a difference between drugs which have the same active substance, but come in different strengths, packages and can be from different manufacturers. So, in order to support a use-case scenario in which a user would be interested in drugs similar to the one he or she is looking for, we decided to create a 'similarTo' relation between each two drugs from our dataset which have the same first seven digits in the ATC code. The relation is defined as both transitive and symmetric in the ontology, which allows more flexibility in the process of querying the data.

In order to transform our drug data into five-star Linked Open Data, we needed relations in the RDF graph towards outside entities. For this purpose, we decided to use the DrugBank dataset, which is the largest and the most detailed drug dataset on the Web. Similarly as in the process of interlinking the drugs internally, we used the ATC codes to detect the similarity between the drugs from our dataset and the drugs from DrugBank. For this purpose, we matched the first seven digits from the ten-digit ATC code in our dataset, with the seven-digit ATC code in the DrugBank dataset. Once the drugs were matched, we added new triples within our graph, denoting that the drug defined in our dataset had an 'owl:seeAlso' relation to the drug defined in the DrugBank dataset. This relation provides new possibilities for data querying, since we can now move from our local drug dataset and get information which is not present locally, but somewhere on the Web, in the LOD and LODD clouds.

E. Publishing the Linked Open Data

Once we had a graph of Linked Open Data from the Health Insurance Fund of Macedonia, the next step was to publish the data on the Web. For this purpose, we created a public instance[12] of Virtuoso at the Faculty of Computer Science and Engineering, in Skopje. This Virtuoso instance holds the Linked Drug Data from the HIFM graph, and provides a public interface via its SPARQL endpoint[13]. The endpoint can be used for querying the drug data from the graph, either by using the SPARQL editor available at the endpoint, or by using the endpoint as a web service from a mobile, web or desktop application.

Additionally, we made dumps of the HIFM graph data into RDF files, represented in RDF/XML, Turtle, N3, RDF/JSON and JSON-LD semantic data formats. These RDF dumps are published on a public CKAN instance[14] at the Faculty of Computer Science and Engineering, in Skopje, from where they can be freely downloaded.

IV. USE-CASES

The purpose of using Linked Open Data is the ability of leveraging the value and usability of the data, in various use-cases. Once we have the local HIFM drug data interlinked with data from the LODD cloud, we can start querying the local data and continue moving through the links to information published elsewhere on the Web. This ability broadens the usage possibilities of the data, and allows development of new types of applications over the data.

A. Using Information from HIFM

Once such use-case would be to use the 'hifm:similarTo' relation to retrieve information about drugs which have the same active substance as the drug we are interested in, but may have a different brand name, different price, may be manufactured by a different company, and may have a different package form and strength.

For instance, if we are looking at information about the drug "NIFADIL, film coated tablets, 50 x 10mg" from the

[12] http://linkeddata.finki.ukim.mk/
[13] http://linkeddata.finki.ukim.mk/sparql
[14] http://data.finki.ukim.mk/

HIFM graph, and we want to find out the drugs which are similar to it, we can use the following SPARQL query:

```
PREFIX drugbank: <http://wifo5-04.informatik.uni-
                mannheim.de/drugbank/resource/drugbank/>
PREFIX hifm: <http://www.fzo.org.mk/ontology/hifm#>

SELECT ?bn ?p ?m
WHERE
{
  hifm:79588 hifm:similarTo ?dbd .
  ?dbd drugbank:brandName ?bn ;
      hifm:refPriceWithVAT ?p ;
      hifm:manufacturer ?m .
}
ORDER BY ASC (?bn)
```

The query first makes a lookup for RDF triples in the HIFM graph where the subject is the drug we are currently interested in, and it is in a 'hifm:similarTo' relation with another drug from the HIFM graph. The drugs similar to the drug with ID = 79588 are placed in the ?dbd variable. Then, we look up the details for these drugs, and select their brand name, the price and the manufacturer (Table 1).

Table 1. Results from the SPARQL query.

Brand Name	*Price*	*Manufacturer*
CORDIPIN R, 30 x 20mg	14,00	KRKA
CORDIPIN XL, 20 x 40mg	19,00	KRKA
KORINCARE NEO, 20 x 40mg	19,00	TCHAIKAPHARMA
KORINCARE, 20 x 20mg	9,00	TCHAIKAPHARMA
NIFADIL RETARD, 30 x 20mg	14,00	ALKALOID
NIFEDIPIN RETARD, 30 x 20mg	14,00	REPLEKFARM
NIFEDIPIN, 50 x 10mg	35,00	JAKA 80
NIFELAT RETARD, 30 x 20mg	14,00	ZDRAVLJE

This query can be written and executed directly in the SPARQL editor at our Virtuoso SPARQL endpoint, or can be sent as a query string from an application, and used as a web service. The web service calls have the following format:

http://linkeddata.finki.ukim.mk/sparql?default-graph-uri=**DEFAULTGRAPH**&query=**SPARQLQUERY**&format=**FORMAT**

Here, DEFAULTGRAPH represents the graph URI of the default graph for the query, i.e. the graph the query should be executed over, SPARQLQUERY represents the SPARQL query, as the one shown above, and FORMAT represents the format of the response, which can be HTML, XML, JSON, Javascript, CSV, Spreadsheet, RDF/XML, N3, Turtle, etc.

With this, a developer of an mobile application over the Linked Open Data from the Fund could easily develop a functionality which, based on the current drug the user is browsing, could offer him alternative drugs which may be more accessible, easier to find, and even cheaper. This would provide the end-user of the application with a better insight into his options as a patient when buying drugs.

B. Using Information from DrugBank, LOD and LODD

Now that the HIFM graph contains links to another dataset on the Web, we can use them to traverse the remote graph. This way, by using the 'owl:seeAlso' relation, we can retrieve information from the DrugBank dataset which are not present in the local HIFM drug data.

For instance, if we want to get information about the food interactions of the drug "DILACOR, tablets, 20 x 0,25mg", we can use the following SPARQL query:

```
PREFIX hifm: <http://www.fzo.org.mk/ontology/hifm#>
PREFIX drugbank: <http://wifo5-04.informatik.uni-
                mannheim.de/drugbank/resource/drugbank/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT ?fi
WHERE
{
  hifm:32964 owl:seeAlso ?dbd
  SERVICE <http://wifo5-04.informatik.uni-
                                mannheim.de/drugbank/sparql>
  {
    ?dbd drugbank:foodInteraction ?fi .
  }
}
ORDER BY ASC (?fi)
```

This SPARQL query starts from the HIFM graph, looking for all of the triples which state that the drug with ID = 32964 is in a 'owl:seeAlso' relation with another drug. The drugs from the matched triples are selected as an ?dbd SPARQL variable, which is then used in the next line and is sent to the SPARQL endpoint at DrugBank. This line asks for triples which will tell us the food interaction for the drug(s) represented by the ?dbd variable. The resulting food interactions will be returned in the ?fi variable, which is then displayed as a result (Table 2).

This information was not stored in our local HIFM dataset, but because of the Linked Open Data principles and the links we provided to drugs published at DrugBank, we were able to retrieve additional information for the given drug. We can use

these types of queries for retrieving any other information which DrugBank provides, for our drugs, defined in our HIFM graph.

Table 2. Results from the SPARQL query.

Food Interactions
Avoid avocado.
Avoid bran and high fiber foods within 2 hours of taking this medication.
Avoid excess salt/sodium unless otherwise instructed by your physician.
Avoid milk, calcium containing dairy products, iron, antacids, or aluminium salts 2 hours before or 6 hours after using antacids while on this medication.
Avoid salt substitutes containing potassium.
Limit garlic, ginger, gingko, and horse chestnut.

This use-case can provide a developer of a medical mobile, web-based or desktop application, with a functionality which would give its end-users additional and vital information about the drugs they are browsing.

The DrugBank dataset contains links to other datasets as well (Fig. 1), and we can use them for accessing other LOD and LODD cloud datasets. For instance, the DrugBank data contain 'owl:sameAs' relations to drugs which are described as part of clinical trials in the LinkedCT dataset. In the same manner as we leap from our HIFM graph to the DrugBank graph, we can continue over to the LinkedCT graph, and gather the needed information from there.

V. Conclusion and Future Work

In this paper, we described the process of transforming the two-star drug data from the Health Insurance Fund of Macedonia into five-star Linked Open Data connected to the DrugBank dataset and the LOD and LODD clouds. The result of the transformation is a HIFM RDF graph, which contains over 21.000 RDF triples for 1.020 drugs from the Fund. These drugs are interconnected with 9.946 'hifm:similarTo' relations with each other, and with 1.015 'owl:seeAlso' relations to drugs from the DrugBank dataset. The HIFM graph is available for use at the live Virtuoso instance and as an RDF dump on the CKAN instance at our Faculty.

We also provided use-cases which give examples of how the data from the Health Insurance Fund and DrugBank can be used, in order to provide application developers with mechanisms and ideas for retrieving distributed data in various formats.

Acknowledgement

The work in this paper was partially financed by the Faculty of Computer Science and Engineering, at the Ss. Cyril and Methodius University in Skopje.

References

[1] T. Berners-Lee, N. Shadbolt, "There's gold to be mined from all our data", *The Times*, 2012.

[2] V. Kundra, "Digital Fuel of the 21st Century: Innovation through Open Data and the Network Effect", *Joan Shorenstein Center on the Press, Politics and Public Policy, Harvard College*, 2012.

[3] C. Bizer, T. Heath and T. Berners-Lee, "Linked Data - The Story So Far" *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009, pp: 1-22.

[4] T. Berners-Lee, *Linked Data - Design Issues*. http://www.w3.org/designissues/linkeddata.html.

[5] R. Cyganiak and A. Jentzsch. *Linking Open Data cloud diagram*. http://lod-cloud.net/.

[6] A. Jentzsch, J. Zhao, O. Hassanzadeh, K. H. Cheung, M. Samwald and B. Andersson, "Linking Open Drug Data", *Triplification Challenge of the International Conference on Semantic Systems*. 2009.

[7] O. Hassanzadeh, A. Kementsietsidis, L. Lim, R. J. Miller and M. Wang, "LinkedCT: A Linked Data Space for Clinical Trials.", *arXiv:0908.0567*, 2009.

[8] K. H. Cheung, E. Prud'hommeaux, Y. Wang and S. Stephens, "Semantic Web for Health Care and Life Sciences: a review of the state of the art." *Briefings in Bioinformatics 10*, 2009, no. 2, pp. 111-113.

[9] M. Mitrevski, M. Jovanovik, R. Stojanov, D. Trajanov, "Open University Data", in *Proceeding from the 9th Conference for Informatics and Information Technology*, 2012.

[10] D. Temelkovski, M. Jovanovik, I. Mishkovski, D. Trajanov, "Towards Open Data in Macedonia: Crime Map based on Ministry of Internal Affairs' Bulletins", in *Proceeding from the 9th Conference for Informatics and Information Technology*, 2012.