Publishing and Interlinking the Global Health Observatory Dataset

Towards a data warehouse for monitoring human development

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Amrapali Zaveri [a,*], Jens Lehmann [a], Sören Auer [a], Mofeed M. Hassan [a], Mohamed A. Sherif [a] and Michael Martin [a]

[a] *Universität Leipzig, Institut für Informatik, AKSW, Postfach 100920, D-04009 Leipzig, Germany*
E-mail: {lastname}@informatik.uni-leipzig.de

Abstract. The improvement of public health is one of the main indicators for societal progress. Statistical data for monitoring public health is highly relevant for a number of sectors, such as research (e.g. in the life sciences or economy), policy making, health care, pharmaceutical industry, insurances etc. Such data is meanwhile available even on a global scale, e.g. in the Global Health Observatory (GHO) of the United Nations's World Health Organization (WHO). GHO comprises more than 50 different datasets, it covers all 198 WHO member countries and is updated as more recent or revised data becomes available or when there are changes to the methodology being used. However, this data is only accessible via complex spreadsheets and, therefore, queries over the 50 different datasets as well as combinations with other datasets are very tedious and require a significant amount of manual work. By making the data available as RDF, we lower the barrier for data re-use and integration. In this article, we describe the conversion and publication process as well as use cases, which can be implemented using the GHO data.

Keywords: RDF, Health, WHO, GHO, statistics, datacube, CSV, spreadsheets

1. Introduction

The improvement of public health is one of the main indicators for societal progress. The *World Health Organization* (WHO)[1], a specialized agency of the United Nations, is mainly concerned with international public health with the main aim of the attainment of the highest possible level of health by all people. Besides publishing reports on global health problems, WHO also provides access to enormous amounts of statistical data and analyses for monitoring the global health situation. The WHO's *Global Health Observatory* (GHO) publishes such statistical data and analyses for important health problems, which is categorised by either country, indicator or topic. The aim of GHO is to provide access to (1) country data and statistics with a focus on comparable estimates, and (2) WHO's analyses to monitor global, regional and country situation and trends[2].

GHO provides access to a wide variety of over 50 different datasets, such as the world health statistics, mortality and burden of disease, health expenditure per capita, deaths due to particular diseases such as HIV/AIDS, Tuberculosis, neglected tropical diseases, violence and injuries, health equity, just to name a few. Each dataset contains an extensive list of indicators

[*] Corresponding author. E-mail: zaveri@informatik.uni-leipzig.de.

[1] `http://www.who.int/en/`

[2] `http://www.who.int/gho/about/en/`

which capture statistical data according to a region, country or based on gender. The data covers all the 198 WHO member countries[3] and while some indicators are from the late 1970s onwards, some are prior to the mid-1990s. The data is updated as more recent or revised data becomes available or when there are changes to the methodology being used. A list of all the datasets with a description of its contents is provided in Table 1.

In this paper, we first describe the process of the conversion of the GHO data to RDF in Section 2. Details of the publishing and the interlinking GHO with other datasets are presented in Section 3. Section 4 portrays a few potential application scenarios and use cases for the GHO data. A number of related initiatives and how GHO is different than what already exists is discussed in Section 5. Finally, we conclude with the lessons learned in Section 6.

2. Dataset Conversion

The GHO data is published as spreadsheets describing a single data item (e.g. death, DALY) in several dimensions (e.g. country, population, disease). In order to convert the data to RDF, we used the RDF Data Cube Vocabulary [2] which is based on the popular SDMX standard[4] and designed particularly to represent multidimensional statistical data using RDF. The vocabulary also uses the SDMX feature of content oriented guidelines (COG). COG defines a set of common statistical concepts and associated code lists that can be re-used across datasets.

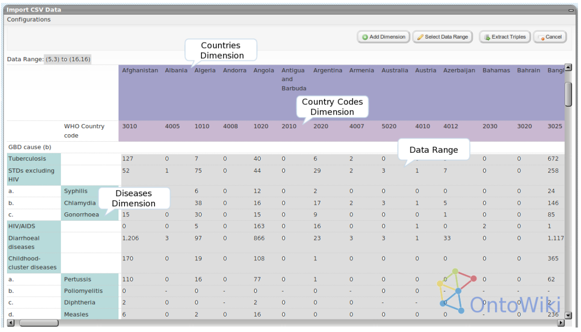
However, transforming these spreadsheets to RDF in a fully automated way may cause information loss as there may be dimensions encoded in the heading or label of a sheet. Thus, we implemented a semi-automatic approach by integrating the algorithm as a plug-in extension in OntoWiki [1]. OntoWiki is a tool which supports agile, distributed knowledge engineering scenarios. Moreover, it provides ontology evolution functionality, which can be used to further transform the newly converted statistical data.

Using this plug-in, when a spreadsheet containing multi-dimensional statistical data is imported into OntoWiki, it is presented as a table as shown in Figure 1. Subsequently, the user has to manually configure the (1) dimensions, (2) attributes by creating them individually and selecting all elements belonging to a certain dimension and (3) the range of statistical items that are measured. Using RDFa, the corresponding COG concepts are automatically suggested, when a user enters a word in the text box provided. The specified configurations can also be saved as a template and reused for similar spreadsheets, such as for data published in consecutive years. Then the plug-in automatically transforms the data into RDF. A presentation detailing the conversion process is available[5].

After converting the GHO data, an RDF dataset containing almost 8 million triples (number of triples for each individual dataset is reported in Table 1) was obtained and published at: `http://gho.aksw.org/`. The mortality and burden of disease dataset in GHO alone accounts for 3 million triples. An example of the death value *127* represented as RDF using the Data Cube vocabulary is illustrated in the following listing:

```
gho:Country  rdfs:subClassOf qb:DimensionProperty;
             rdf:type rdfs:Class;
             rdfs:label "Country" .

gho:Disease  rdfs:subClassOf qb:DimensionProperty;
             rdf:type rdfs:Class;
             rdfs:label "Disease" .

gho: Afghanistan  rdf:type ex:Country;
                  rdfs:label "Afghanistan" .

gho:Tuberculosis  rdf:type ex:Disease;
                  rdfs:label "Tuberculosis" .

gho:c1-r6    rdf:type      qb:Observation;
             rdf:value     "127"^^xsd:integer;
             qb:dimension  gho:Afghanistan;
             qb:dimension  gho:Tuberculosis .
```

Listing 1: RDF representation of the death value ’127’ using the RDF Data Cube Vocabulary



Data Range: (5,3) to (16,16)

		Afghanistan	Albania	Algeria	Andorra	Angola	Antigua and Barbuda	Argentina	Armenia	Australia	Austria	Azerbaijan	Bahamas	Bahrain	Bangl
	WHO Country code	3010	4005	1010	4008	1020	2010	2020	4007	5020	4010	4012	2030	3020	3025
GBD cause (b)															
Tuberculosis		127	0	7	0	40	0	6	2	0	[illegible]	[illegible]	0	0	672
STDs excluding HIV		52	1	75	0	44	0	29	2	3	1	7	0	0	258
a.	Syphilis	[illegible]	[illegible]	6	0	12	0	2	0	0	0	0	0	0	24
b.	Chlamydia	[illegible]	[illegible]	38	0	16	0	17	2	3	1	5	0	0	146
c.	Gonorrhoea	15	0	30	0	15	0	9	0	0	0	1	0	0	85
HIV/AIDS		0	0	5	0	163	0	16	0	0	1	0	2	0	1
Diarrhoeal diseases		1,206	3	97	0	866	0	23	3	3	1	33	0	0	1,117
Childhood-cluster diseases		170	0	19	0	108	0	1	0	0	0	0	0	0	365
a.	Pertussis	110	0	16	0	77	0	1	0	0	0	0	0	0	62
b.	Poliomyelitis	0	-	0	-	0	-	0	-	0	0	-	-	-	-
c.	Diphtheria	2	0	0	-	2	0	0	0	-	-	0	0	0	2
d.	Measles	6	0	2	0	16	0	0	0	0	0	0	-	0	236

Fig. 1. Screenshot of the OntoWiki statistical data import wizard displaying a GHO table configured for conversion into RDF.

[3] `http://www.who.int/countries/en/`

[4] `http://sdmx.org`

Table 1

Different statistical datasets available in the Global Health Observatory.

Dataset	**Description**	**No.of triples**
Environmental health	Number of deaths due to children health, climate change, household air pollution, UV radiation, water, sanitation and hygiene	31,012
Epidemic prone diseases	Number of reported cases of cholera, meningococcal meningitis and statistics from the Global Influenza Surveillance and Response System	255,957
Equity	Equity figures for women health, urban health and social determinants of health	324,445
Health-related Millennium Development Goals	Health indicators associated with poverty and hunger, child mortality, maternal health, environment sustainability, and global partnership for development.	784,346
Health systems	Data on healthcare infrastructure, essential health technologies, aid effectiveness, health financing, essential medicines, service delivery and health workforce	234,340
HIV/AIDS	Data on the size of the epidemic and on the HIV/AIDS response	99,476
Immunization	Country and regional data of immunisation efforts for several diseases	625,082
Injuries and violence	Number of deaths due to road traffic accidents, data on demographic and socio-economic statistics, emergency care provision and existence of a national policy for human safety	242845
Mortality and burden of disease	Number of deaths, DALYs, life expectancy, mortality and morbidity, disease and injury country estimates for each country	3,000,000
Neglected Tropical Diseases	Statistics on newly reported cases of each of the neglected tropical disease that is monitored	167,841
Noncommunicable Diseases	Mortality measures, risk factors and health system response and capacity for each of the noncommunicable disease that is monitored	1,409,629
Tobacco Control	Data on the prevalence of adult and youth consuming tobacco and various measures to help prevent tobacco consumption, such as policies, help, warnings, enforcing bans	379,283
Tuberculosis	Cases of incidence and mortality, diagnosis, drug regimens, treatment success for tuberculosis in each country	67,479

Table 2

Technical details of the GHO RDF dataset.

URL	http://gho.aksw.org/
Version date and number	01-11-2010, 1.0
Licensing	WHO allows reproduction of its data for non-commercial purposes.
VoiD File	`http://db0.aksw.org/downloads/void.ttl`
The DataHub entry	`http://thedatahub.org/dataset/gho`

3. Dataset Publishing and Linking

Dataset Publishing. After converting the GHO data as RDF, we published it as Linked Data using the *OntoWiki* platform [1]. OntoWiki not only allows the publishing and maintenance of the data but also provides a SPARQL endpoint for the dataset in combination with *Virtuoso*[6] as the storage solution for the RDF model. Additionally, it is also possible to browse the data with the HTML output of OntoWiki. Details and links of the SPARQL endpoint, the version, licensing, availability and link to the VoiD file are listed in Table 2.

Listing 2 provides an example of a SPARQL query which retrieves the number of deaths (from GHO) and the number of clinical trials (from LinkedCT) for the disease Tuberculosis and HIV/AIDS in all countries.

Dataset Interlinking. We used the *mortality and burden of disease dataset* from GHO as a test-environment for link generation and linked it with the *LinkedCT*[7] (the Linked Data version of ClinicalTrials.gov) and *PubMed*[8] (converted to Linked Data by the Bio2RDF project) datasets. We used the *Silk 2.0* [3] tool, which is developed for discovering relationships between data items within different knowledge bases, that are avail-

[5] `http://goo.gl/OHDM9`

[6] `http://virtuoso.openlinksw.com/`

[7] `http://linkedct.org/`

[8] `http://bio2rdf.org/`

able via SPARQL endpoints. Silk includes a declarative language for specifying (1) the types of RDF links that should be discovered and (2) the conditions which the data items must fulfill in order to be interlinked. We used the *Jaro distance* as string metric where applicable and two confidence value thresholds: (1) Links above 0.95 confidence were accepted and (2) links between 0.90 and 0.95 were saved to a separate file for manual inspection. The number of interlinks obtained for countries and diseases is displayed in Table 3.

```
SELECT ?countryname ?diseasename ?value
count(?trial)
FROM <http://gho.aksw.org/>
FROM <http://linkedct.org/>
WHERE {?item       a           qb:Observation ;
              gho:country      ?country ;
              gho:disease      ?disease ;
              att:unitMeasure  gho:Measure ;
              gho:incidence     ?value .
  ?country    rdfs:label       ?countryname .
  ?disease    rdfs:label       ?diseasename .
  ?trial      a                ct:trials ;
              ct:condition     ?condition ;
              ct:location      ?location .
  ?condition  owl:sameAs       ?disease .
  ?location   shv:locatedIn    ?country .
FILTER (?diseasename("Tuberculosis", "HIV/AIDS")).
} GROUP BY ?countryname ?diseasename ?incidence
```

Listing 2: SPARQL query for retrieving the number of deaths and number of trials for Tuberculosis and HIV/AIDS in all countries.

In addition to the ability to explore the data by using SPARQL and the resulting lists of resources, users are able to visualize the data by using *CubeViz*. CubeViz is an OntoWiki extension, which uses DataCube resources as input. After the selection of desired dimension properties such as `gho:disease` and `gho:country` as well as the measure property `gho:incidence` CubeViz is able to generate different type of charts (e.g. bar chart, pie chart, spline chart)). As an example, the incidence of the disease "Migraine" in selected countries can be visualized with CubeViz as depicted in Figure 2.

4. Application Scenarios and Use-Cases

In this section, we outline selected application scenarios and use-cases for the Linked GHO data.

Monitoring health care scenarios. Since GHO provides information on mortality, morbidity, health status, service coverage and risk factors for several diseases in each country, it can be used by each country to

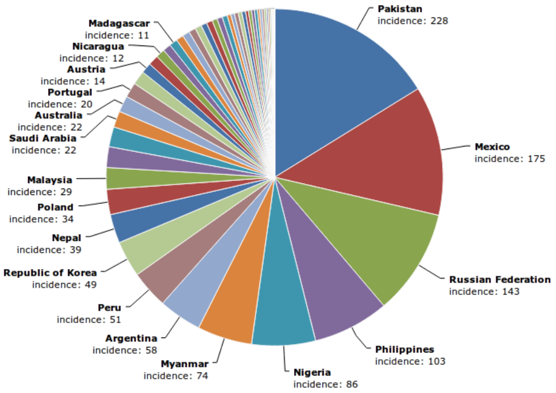


Fig. 2. Screenshot of CubeViz displaying the pie chart of incidence of Migraine in a subset of countries.

monitor the disease prevalence for any given year and to compare prevalence as well as the effect of countermeasures with similar or neighboring countries. This can help to implement either precautionary measures if the mortality is high or curb health expenditures for diseases which seem to have adequate treatment options.

Disparity Analysis. Another application of the GHO dataset is evaluating the disparity between the availability of treatment options and the global burden of disease, as illustrated in the ReDD-Observatory project [4]. This disparity is partially caused by the limited access to information that would allow health care and research policy makers make more informed decisions regarding health care services. The hindrance lies in reliably obtaining and integrating data regarding the disease burden and the respective research investments. Therefore, as the Linked Data paradigm provides a simple mechanism for publishing and interlinking structured information on the Web, an opportunity is created to reduce this information gap that would allow for better policies in response to these disparities.

Primary source providing ground truth. GHO enables direct linking to the ground truth data for secondary (e.g. scientific publications) or tertiary (e.g. encyclopedias) sources. This enables improved provenance tracking in those sources. It also allows automatic syndication of the data using SPARQL or simple REST queries, which enables a simpler verification of statements compared to the manual work, which would be necessary without Linked Data.

Human development data warehouse. Just as data warehouses and business intelligence are now inte-

Table 3

Number of links obtained along with precision values between GHO, PubMed and LinkedCT for diseases and countries.

Link		Source		Target		Links		Precision	
between	type	Dataset	Instances	Dataset	Instances	Accepted	To verify	Accepted	To verify
Diseases	owl:sameAs	LinkedCT	5000	GHO	128	163	43	0.9625	-
Diseases	rdfs:subClassOf	LinkedCT	5000	GHO	128	469	45	1	0.9926
Diseases	owl:sameAs	GHO	128	PubMed	23618	453	75	1	0.7083
Locations	redd:locatedIn	LinkedCT	757341	GHO	192	300000	0	1	-
Countries	owl:sameAs	GHO	192	PubMed	23618	201	12	1	0.9583

gral parts of every larger enterprise, the linked GHO data can be the nucleus for a human development data warehouse. In such a human development data warehouse, a large number of statistical data and indicators are published by different organizations that could be integrated automatically or semi-automatically in order to obtain a more interactive picture of the human development. Currently, the indicators (e.g. the Human Development Index) are very coarse-grained, mainly referring to countries. Using linked data, such indicators could be computed on a much more fine-grained level, such as for cities and regions as well as with regard to different groups of people (e.g. per gender, ethnicity, education level). Policy making would be based on more rational,transparent and observable decisions as it is advocated by evidence-based policy.

5. Related Initiatives

There are already a number of efforts to convert health care and life science related data sets to Linked Data such as LODD, LinkedCT, OBO ontologies and the W3C's Health Care and Life Sciences Working Group, each of which is discussed in this section along with the importance of converting and publishing the GHO datasets.

LODD, i.e. the Linking Open Drug Data project[9], mainly converts, publishes and interlinks drug data that is available on the web, ranging from impacts of drugs on gene expression to results of the clinical trials. A number of datasets have been converted in this project[10] including DrugBank, DailyMed, SIDER to name a few. However, these datasets are restricted to drug data and even though they do contain disease data (from the Diseasome dataset), they do not connect the number of deaths or the health expenditure or the status of the health system in each country for each of the diseases that are included (as provided by GHO).

LinkedCT is the Linked Data version of Clinical-Trails.gov which publishes data about clinical trials in RDF and links it to other datasets such as PubMed. Even though, in LinkedCT each trial is associated with a disease and a drug, it does not provide information about the prevalence of the disease in a particular country, which is provided in GHO.

OBO is the Open Biological and Biomedical Ontologies project[11] which aims to create a suite of interoperable reference ontologies in the biomedical domain. It brings together biology researchers and ontology developers who work together to develop a set of ontologies as well as design principles that can help develop interoperable ontologies. However, most of the ontologies developed are at the experimental level or organismal level and are not yet sufficiently interlinked with other datasets available as Linked Data.

The Semantic Web Health Care and Life Sciences (HCLS Interest Group)[12] is established by the World Wide Web Consortium (W3C) to support the use of Semantic Web technologies in health care, life sciences, clinical research and translational medicines. The group focuses on aiding decision-making in clinical research, applying the strengths of Semantic Web technologies to unify the collection of data for the purpose of both primary care (electronic medical records) and clinical research (patient recruitment, study management, outcomes-based longitudinal analysis, etc.). Subgroups, on the other hand, focus on making the biomedical data available in RDF, dealing

[9] http://www.w3.org/wiki/HCLSIG/LODD/

[10] http://www.w3.org/wiki/HCLSIG/LODD/Data

[11] http://obofoundry.org/

[12] http://www.w3.org/blog/hcls/

with biomedical ontologies, focus on drug safety and efficacy communication and support researchers in the navigation and annotation of the large amount of potentially relevant literature.

6. Conclusion and Lessons Learned

Although we were able to successfully convert the GHO dataset and utilize one of the datasets in a use case, we encountered some problems such as cumbersome conversion, low interlinking quality and lack of time series capability in the datasets. We discuss these problems in the sequel.

Conversion. The conversion process was time consuming and cumbersome because, first of all, each individual Excel files needed to be downloaded from the GHO web portal. Each file had to be then converted into CSV so that it could be appropriately displayed as an HTML table in OntoWiki. Since the conversion method was semi-automated, one had to individually selected the dimensions, attributes and data range for each of the files. While some of the required steps such as the annotation of the CSV files for conversion are not automatizable, other steps, such as the Excel to CSV conversion can be performed more efficiently (e.g. in a batch or through bulk processing).

Coherence. The number of interlinks obtained between the datasets for diseases was relatively low, as presented in Table 3. The main reason was the different use of identifiers for the naming of the disease. For example, 'heart attack' in GHO could not be matched with 'cardiac arrest' in LinkedCT using the basic string similar functionality of SILK. In order to address this problem, we plan to extend current linking tools in such a way, that background knowledge in the form of gazetteers can be also taken into account.

Temporal Comparability. The data in GHO is not published regularly every year. Also, since the health data recording and handling systems differ between countries, comparability of the data is limited. This is mainly due to the differences in definitions and/or time periods and incomplete data provision from different countries. Therefore, computing time trends is not possible using GHO, which would be a good indicator of the health scenario in each country over a number of years. We expect, however, that the increased visibility and transparency of a Linked Data version of GHO together with the enhanced possibility of annotation and linking (when compared to simple Excel sheets) will contribute to standardization and increased temporal comparability in the future.

Exploring GHO. Using OntoWiki or similar tools (such as Disco or Tabulator etc.) to browse the RDF data helps users to gain new insights. CubeViz as an OntoWiki extension provides visualization of statistical data (such as GHO) in an user friendly way by means of displaying the data in various types of diagrams and charts. However, a limitation of such generic visualization tools is their limited scalability.

Conclusion. In conclusion, by providing the GHO data as Linked Data and interlinking it with other datasets, it is possible to not only obtain information on important health related topics in each country but also ease the work of health care professionals for data analysis in providing easy access to data. Moreover, it provide opportunities to link to related data and thus perform analyses for current priority health issues.

The Linked Data publishing and linking of the GHO data is a first milestone in a larger research and development agenda: The creation of a global human development data warehouse, which allows to interactively monitor social, societal and economic progress on a global scale.

7. Acknowledgment

This work was supported by grants from the European Union's 7th Framework Programme provided for the projects LOD2 (GA no. 257943) and LATC (GA no. 256975). We thank the LATC and LOD2 consortium members as well as the European Commission for their feedback on the GHO dataset.

References

[1] S. Auer, S. Dietzold, and T. Riechert. Ontowiki - a tool for social, semantic collaboration. In *ISWC 2006*, volume 4273 of *LNCS*. Springer, 2006.

[2] J. Tennison, R. Cyganiak, and D. Reynolds. The rdf data cube vocabulary. Technical report, W3C, 8 2012. http://www.w3.org/TR/vocab-data-cube/.

[3] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Discovering and maintaining links on the web of data. In *ISWC*, 2009.

[4] A. Zaveri, R. Pietrobon, S. Auer, J. Lehmann, M. Martin, and T. Ermilov. Redd-observatory: Using the web of data for evaluating the research-disease disparity. In *2011 IEEE / WIC / ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2011.