COVIDCube: An RDF Data Cube for Exploring Among-Country COVID-19 Correlations

Tamara Novoa-Rodríguez and Aidan Hogan

DIE/DCC, University of Chile; IMFD; Santiago, Chile;
tamara.novoa@ug.uchile.cl, ahogan@dcc.uchile.cl

Abstract. We present an RDF Data Cube – integrated from numerous sources on the Web – that describes countries in terms of general variables (e.g., GDP, population density) and COVID-19 variables. On top of this data cube, we develop a system that computes and visualises correlations between these variables, providing insights into the factors that correlate with COVID-19 cases, deaths, etc., on an international level.
Demo link: https://c19.dcc.uchile.cl/

1 Introduction

The among-country variation in the numbers of reported COVID-19 cases and deaths per capita is not well-understood [6]. Various hypotheses have been proposed, such as high prevalence of comorbidities, climate, pollution, health services, public policies, etc. While a number of initiatives have explored this variance (e.g., [6]), or have developed relevant datasets to better understand this variance (e.g., [4]), there remains uncertainty regarding the factors involved.

In this demo, we will discuss on-going work regarding the preparation of an RDF Data Cube that collects together a wide variety of both general and COVID-19-specific variables at a country level. On top of this RDF Data Cube, we have built a system to visually explore the correlations that exist between such variables in order to gain insights on the among-country variations observed for COVID-19. We call this data cube and associated system "COVIDCube". Code is made available online at https://github.com/tmnvrd/COVIDCube.

2 COVIDCube

We first describe the data preparation for generating the RDF Data Cube. Thereafter we describe the system used to visualise correlations.

RDF Data Cube: There are a wide range of datasets available online that provide different types of indicators for countries. In order to narrow down the scope of the variables considered, we conducted an initial survey in search of hypotheses relating to the among-country variations observed for reported COVID-19 cases and deaths. From these, we broadly identified three main categories of indicators relating to *economics* (e.g., GDP, wages, unemployment), *health* (e.g., obesity and other comorbidities, blood type, vitamin-C deficiency, adult and child mortality rates), and *climate* (e.g., temperature, precipitation, pollution). We also identified hypotheses relating to *miscellaneous* factors, such as political ideologies, transportation networks, tourism, etc. Based on these factors, we began to identify potential sources of data online at the international level, extracting 525 variables from sources including Our World In Data, Wikipedia, The World Health Organization, The World Bank, among others. These datasets, mostly tabular in nature, were extracted as CSV files. We further extracted data for 4 variables pertaining to COVID-19 at the international level from Johns Hopkins University Center for Systems Science and Engineering (CSSE) and Our World in Data (OwID), namely: confirmed cases (CSSE), confirmed deaths (CSSE), confirmed recovered (CSSE), and stringency index (OWiD)[1]

The data were diverse: some datasets were broken down by temporal or regional dimensions; measures were provided in different units; naming variations were present for countries or regions; etc. We wished to integrate the data while modelling their provenance. We thus chose to adopt the RDF Data Cube vocabulary and model [1], using handcrafted Tarql mappings[2] to convert from the raw CSV for each variable to the desired RDF data. Each variable was treated as a distinct dataset and we manually added metadata (using PROV-O and other standard vocabularies) to allow for tracking provenance. Entities mentioned in the datasets – relating to countries, regions, types of disease, genders, education level, etc. – were mapped to their corresponding Wikidata identifiers, thus resolving variations in naming across datasets. Given the time-consuming nature of this task, rather than mapping all 529 variables to a dataset, we identified and converted 79 variables specifically relating to hypotheses found during our survey (including the 4 COVID-19 variables), along with an additional 39 datasets for other variables of interest, resulting in an RDF Data Cube containing 118 different datasets/variables integrated from eight distinct sources, with a total of 442,420 individual observations covering 251 countries (or territories). The data are hosted in a Fuseki SPARQL endpoint.[3]

Visualisation: In order to visually explore among-country correlations between COVID-19 variables and other variables, we built a prototype system in Flask – a micro web framework written in Python – to query the Fuseki back-end for the available pairs of variables, and compute correlations for them. In terms of

[1] The stringency index is a composite measure used to indicate the strictness of public policies to curtail COVID-19 transmission, including travel bans, school closures, etc.
[2] `https://tarql.github.io/`
[3] See `https://c19.dcc.uchile.cl/db/dataset.html?tab=query&ds=/ds`.

Fig. 1. Heat-map with most positive correlations (Spearman's ρ) for health variables

the correlation measures, we chose to use Pearson's r and Spearman's ρ, both of which provide a value in the interval $[-1, 1]$, with -1 indicating perfect negative correlation, 0 no correlation, and 1 perfect positive correlation. We further calculate p-values in order to indicate the probability of the null hypothesis given the observed variables: namely that there is no relation between the variables. We noticed that a confounding factor for many of the variables presented related to population: we thus normalised selected variables by population in the query prior to calculating the correlations. The results are then visualised in four heat-map matrices, collecting together variables categorised by *economics*, *health*, *climate* and *miscellaneous*. Each matrix has rows denoting general variables and columns denoting the four COVID-19 variables, with the rows ordered from highest correlation to lowest correlation with respect to the total number of COVID-19 cases. To improve response times, data are cached in the front-end.

We provide a screenshot of part of the visualisation in Figure 1 referring to the top-10 health variables in terms of positive correlation to confirmed COVID-19 cases. The colours and values indicate the value of correlation, where the colours range from red (positive) to blue (negative). Results that reach a particular level of statistical significance ($p < \alpha$) are noted with $\cdot^*$ ($\alpha = 0.05$) and $\cdot^{**}$ ($\alpha = 0.01$). We see that variables such as life expectancy, body mass index, prevalence of overweight children, obesity, blood types A and O, etc., correlate positively in terms of the number of confirmed COVID-19 cases. The reader may notice that Prevalence of overweight children appears twice in the results. This is because y-axis labels are sometimes hierarchical, where only the first level of the label is shown to avoid overly-long variable names; dimension 5 refers to male children, while dimension 6 refers to female children, which can be seen by hovering over the respective result in the interface. Further such results for other variables can be explored in our online demo: `https://c19.dcc.uchile.cl/`.

Evaluation: In order to initially understand users' opinions of the system, we created a quantitative survey that was published in a university forum and received 52 responses. On a Likert scale of 1–5, users were most positive with respect to the functionality (average: 4.25) and usefulness (average: 4.12) of the platform,

but were less positive regarding how easy it was to understand the information provided (average: 2.98), where users require some base statistical knowledge of correlations, p-values, etc., in order to fully understand the data presented.

3 Discussion

Correlation obviously does not imply causation, but correlations may lead to novel hypotheses and prompt further study regarding potential underlying factors for among-country variations. Though an in-depth analysis of the correlations found is out-of-scope, in summary, we did find a number of variables spanning the different categories that had statistically significant positive or negative correlation with the COVID-19 variables: more than would be expected according to the chosen significance level (e.g., we find more than the expected 1/20 variables satisfying the $\alpha = 0.05$ threshold). Examples of significant correlations can be seen in Figure 1, and also by visiting our demo.

Some correlations observed were to be expected, such as in the case of positive correlations for obesity, which is associated with comorbidities for COVID-19. However, other correlations were surprising. For example, one of the most negatively correlated health-related variables was *No access to handwashing facility*, which suggests, with statistically significant results, that countries where *fewer* people can wash their hands tend to have *fewer* per-capita cases of COVID-19. This seems counter-intuitive as handwashing is considered a way to reduce transmission of the virus. Another (initially) counter-intuitive result is that increased health expenditure per capita correlates positively with COVID-19 cases, where one would expect that better health services would help to reduce transmission and cases. However, by considering the overall matrices, a common confounding factor begins to emerge: namely the level of development of the country, with more developed countries tending to have more confirmed cases. Similar observations can be found elsewhere, and possible explanations include a lack of testing [5] and an increased remoteness [2] of developing countries.

While it is not possible to draw firm conclusions about the among-country variance observed for COVID-19 from the currently available data, COVIDCube does provide some insights and clues as to potentially important factors. Deriving more definitive conclusions will require integrating diverse data from further diverse sources, for which RDF Data Cubes provide a relevant solution.

Regarding the use of RDF Data Cubes, in the original plan for the project, we had intended to use a relational database as our back-end to load different variables by country. However, when accessing the raw data, we encountered a variety of issues relating to their diversity, including the use of different units and multipliers; different names being used; different countries being included/excluded from certain sources; the same demographics (such as gender, age, etc.) appearing in different measures; some countries having their results being presented by state or region; different measures having different temporal granularities; etc. Rather than cleaning and preprocessing the data to "fit" a clean relational schema, we rather chose to use RDF Data Cubes to repre-

sent the diverse underlying data in a more complete way, and thereafter use SPARQL queries to compute the tables from which correlations were extracted. This approach offered a greater decoupling between the data preparation and the application design, where we could focus on representing the underlying data as completely as possible in the RDF Data Cube, and then later use SPARQL queries to extract the data needed for the application from multiple sources.

In terms of future directions, we wish to investigate correlations along temporal dimensions; such data are available in the data cube but are not exploited by the visualisation. Similarly, although the underlying dataset tracks provenance information, this is not shown in the current interface; adding links to the underlying sources used would help others to build upon and reproduce the results shown. Incorporating other data sources – including integration with existing RDF datasets (e.g., on COVID-19 [3,7]) – would also be of interest to enrich the data and enhance the analyses possible. As COVID-19 remains an ongoing phenomenon, it would also be of interest to implement a framework to automatically update the statistics based on the underlying sources.

Acknowledgements We would like to thank Bryan Ortiz Pizarro, Catalina Rojas Zúñiga, Cecilia Pilar Mancill, Clemente Parades Gómez, Cristóbal Masías Durán, José Miguel Pacheco, Loreto Palma Donoso, Osvaldo Garay Roos, Sebastián Aguilera Valenzuela, Tomás Torres Bardavid and Valentían Espina Carmona for their considerable help with mapping raw data from CSV to Turtle. We also thank the anonymous reviewers for their very helpful feedback. This work was supported by ANID – Millennium Science Initiative Program – Code ICN17_002 and by FONDECYT Grant No. 1181896.

References

1. Cyganiak, R., Reynolds, D., Tennison, J.: The RDF Data Cube Vocabulary. W3C Recommendation (2014), `https://www.w3.org/TR/vocab-data-cube/`
2. Gisselquist, R.M., Vaccaro, A.: Why countries best placed to handle the pandemic appear to have fared the worst. The Conversation (2021)
3. Michel, F., et al.: Covid-on-the-Web: Knowledge Graph and Services to Advance COVID-19 Research. In: ISWC. pp. 294–310 (2020)
4. Miller, L.E., Bhattacharyya, R., Miller, A.L.: Data regarding country-specific variability in Covid-19 prevalence, incidence, and case fatality rate. Data in Brief (2020)
5. Nordling, L.: Africa's pandemic puzzle: why so few cases and deaths? Science **369**(6505), 756–757 (2020)
6. Sorci, G., Faivre, B., Morand, S.: Explaining among-country variation in COVID-19 case fatality rate. Scientific Reports **10**(18909), 1493–1500 (2020)
7. Steenwinckel, B., et al.: Facilitating the Analysis of COVID-19 Literature Through a Knowledge Graph. In: ISWC (2020)