COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Peter Rose (UC San Diego), May 5,2020

Title: COVID-19-Net: Integrating Health, Pathogen and Environmental Data into a Knowledge Graph for Case Tracking, Analysis, and Forecasting

Peter Rose CIC Database Profile

NSF Award #: 2028411

Youtube Recording with Slides

July 2020 CIC Webinar Information

Transcript Editor: Rhyley Vaughan

## Transcript

Florence Hudson:

Thanks, Peter.

Peter Rose:

*Slide 1*

Okay. Thank you, Florence, for inviting us. Let me tell you a little bit about the background of COVID-19-Net. We were part of the NSF open knowledge network program. In particular, we were interested in linking three types of data: biomedical data, environmental data, and socio-demographic data. Then, in January COVID came along. We thought this is really a great use case where we need to integrate multidisciplinary data, so if we can go to the next slide.

*Slide 2*

That shows you what we're trying to accomplish, so if you think about COVID, there are really three major areas. One is the host, which can be human or can be animals, and so on. Then, we have the pathogen, the virus, and then everything else, which we call the environment. We want to enable researchers basically to look at this interplay between those different areas. For example, looking at host-pathogen interaction: how does environment affect infections and so forth? That basically was the background, and we started. Take the next slide please.

*Slide 3*

When we applied for this RAPID grant, we had collaborators on the Open Knowledge Network from the NSF. We had collaborators from UC Santa Barbara. It's in green. UCSF is in orange, and us in blue. We kind of divide it up in particular areas. We are focusing on the blue areas like population characteristics, health data, pathogen information, and environmental information, whereas our collaborators from UC Santa Barbara (this is another RAPID) focuses on transportation and supply chain. UCSF focuses on biomedicine. Next slide please.

*Slide 4*

This is, kind of, our prototype knowledge graph. In a knowledge graph you don't have information silos, but you link all the data together. This is just a prototype. We don't have much information in here yet. On the right-hand-side, you see information about the pathogen, the host. We have information about the epidemiology, about this outbreak here. Then we focused a lot on the biomedical area. We have more than 30,000 different strains in this knowledge graph, but it's all linked together. For each strain, we know all its variants of mutations, the effect on the genes and proteins. We know about protein-protein interactions. We link this together with publications. Then, most importantly, we also link this to geolocation. We mapped out the entire geographic hierarchy of the world, so we can map strains or cases to any location in the world - all the way down to the census-tract level. Okay, next slide please.

*Slide 5*

When we started this project, we wanted this to be an automated project that can be expanded by others, so we have a very transparent and reproducible workflow. First of all, we start off with open access data. We want to be able to redistribute the information, so we start with trustworthy public data repositories, and then we created a process that automatically extracts information and integrates that information. That's really the key— the integration of all this information. We spend a lot of [...] this is where most of the work goes. With COVID, things change on a daily basis, so we actually have a daily update process. We have open source software that is in, and on a daily basis, we update information, integrate information, and then upload that into a knowledge graph that then can be queried next. We try to follow the fair principles, and everything is open and easily accessible. You can get to all our software. It's reusable and so forth, so with that, maybe go to our next slide.

*Slide 6*

You know once we create this knowledge graph, there are a few things we can (or the end user) can do. First, you can query and browse the knowledge where I find information. What's shown here on the top left. This is exploring protein-protein interaction between virus proteins and human proteins, for example. Then, since it's in a graph form, you can interact if we actually explore this. In green, are various strains of the virus. In the particular geolocation, you can look at specific mutations and see how they're being shared among different strains, for example. This is a more interactive analysis than an exploratory analysis, but if you want to do a more quantitative in-depth analysis, you can also access all that data easily in computational notebooks, such as our studio or Jupyter notebooks for a more reproducible type of analysis. As I mentioned, mapping onto geolocation is obviously very important for corvettes, so we also access this information through dashboards. Up on the top right, we show, for example, San Diego County current cases or predicted case counts. In the center here, we focus in on specific cities and look at, for example, various pre-existing conditions in those areas like: now what's the prevalence of cancer, diabetes, heart disease, and so forth, and how does it affect the population at risk? We can drill down further using this swell tool. We can drill down, for example, all the way to the census-tract level and then look in more depth at its health populations at risk, age structure, and so forth. We want to thank you. Next slide please.

*Slide 7*

We want to thank the NSF for funding, and also the Open Knowledge Network program we were part of, and the collaborators there. Obviously, we're also looking for collaborators. We're looking for a number of collaborators, so if you have open data sets you want to share with us, we like to discuss them. We like to integrate them. If you have code that extracts data from various data sources that will be of interest to us, and obviously if you want to use any of our data (and I think there we already talked to a number of people in this program), you know we want to hear from you too. I think that's about it, so in the chat window I'll paste a link to our COVID graph so it's all available online, so you can explore that yourself. Thank you.