# Surfing towards open data integration – more information in the webinar 22 September 2021!

Sometimes different professionals need to be able to quickly access the same data and interpret them for very different purposes or from many points of view. One example is the COVID-19 emergency: the same data needs to be accessible and easy to use by virologists, epidemiologists, medical doctors, geneticists, social scientists, and many for professions. Stefano Ceri, professor in Data Management at Politecnico di Milano, and his research group built ViruSurf: a simple user interface that enables searches over their database of 3 million SARS-CoV-2 sequences. Read more and join the upcoming webinar 22 September 2021!

Opening data is just the first of the many steps toward data re-use. Data are submitted to field-specific repositories, with specific data formats and described with the addition of metadata. Metadata are truly the rosetta stone that will make data interpretable for future re-use: they bring all the information together to facilitate secondary use in research and provide ways for researchers around the world to understand how data were collated originally and what all parts of the data meant to the primary study.

Data repositories use a large selection of ontologies, vocabularies and metadata formats which can confuse end-users.

Nowadays, different professionals need to be able to quickly access the same data and interpret them for very different purposes or from many points of view. One example is the COVID-19 emergency: the same data needs to be accessible and easy to use by virologists, epidemiologists, medical doctors, geneticists, social scientists, and many for professions.

Stefano Ceri, professor in Data Management at Politecnico di Milano, and his research group, worked toward a solution for SARS-CoV-2 data sequence integration. They developed architecture and pipelines to integrate metadata obtained from different sources. They built VirusSurf: a simple user interface that enables searches over their database of 3 million SARS-CoV-2 sequences. The integrated and curated database is continuously updated from several repositories: GenBank, COG-UK, and GISAID. VirusSurf allows users to make complex search queries that can provide comprehensive answers, a unique feature compared to existing systems. The tool can be used side by side with user interfaces developed to visualize the search results (VirusViz), raw data analysis (VirusLab) and facilitate specific analysis related to vaccine development (EpiSurf).

ViruSurf was developed based on the input received by interviewing a variety of experts of the various aspects of virus research (including clinicians, epidemiological experts, drug and vaccine developers). "The most unexpected aspect of our interview process was the availability and enthusiasm of the experts that we interviewed." says Anna Bernasconi, postdoctoral researcher of Prof. Ceri's group. "We managed to talk with experts from all fields, they were excited in sharing their knowledge with us and actually available above our expectations."

Prof. Ceri adds "The most difficult aspect we had to consider is the growth of data, which went up from about one hundred thousand genomes when we started up to 3.5 million genomes as of today: they need to be mastered by increasingly powerful computing resources. Moreover, while we learn how to solve problems, we face new challenges – we are developing new tools for data analysis that can be used for detailed tracing of mutation prevalence in time and space, with the objective of helping the analysis and control of the viral spreading."

In addition to working on viral data integration, the research group participated in the COVID-19 Host Genetics Initiative, where they developed a metadata model used to collect and harmonize human phenotypic data by hundreds of researchers worldwide.

If you are interested to know more about ViruSurf and other related research on viral genomics, join the upcoming webinar.

## Webinar

**Speaker: Prof. Stefano Ceri, Politecnico di Milano**

**Title: Storing and Analyzing SARS-CoV-2 Viral Sequences through Data-driven Genomic Computing**

**Time: 22 September 2021 at 5:00pm CET**

**URL:** https://www.informatics-europe.org/activities/webinars.html

Resources:

http://gmql.eu/virusurf/

http://www.bioinformatics.deib.polimi.it/geco/?try_virus

https://bio.tools/virusurf

References:

Arif Canakoglu, Pietro Pinoli, Anna Bernasconi, Tommaso Alfonsi, Damianos P Melidis, Stefano Ceri, ViruSurf: an integrated database to investigate viral sequences, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D817–D824, https://doi.org/10.1093/nar/gkaa846

Anna Bernasconi, Andrea Gulino, Tommaso Alfonsi, Arif Canakoglu, Pietro Pinoli, Anna Sandionigi, Stefano Ceri, VirusViz: comparative analysis and effective visualization of viral nucleotide and amino acid variants, Nucleic Acids Research, 2021; gkab478, https://doi.org/10.1093/nar/gkab478

Anna Bernasconi, Arif Canakoglu, Marco Masseroli, Pietro Pinoli, Stefano Ceri, A review on viral data sources and search systems for perspective mitigation of COVID-19, Briefings in Bioinformatics, Volume 22, Issue 2, March 2021, Pages 664–675, https://doi.org/10.1093/bib/bbaa359

COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. Nature (2021). https://doi.org/10.1038/s41586-021-03767-x

C S C
www.csc.fi

elixir
elixir-europe.org