CoronaWhy: Building a Distributed, Credible and Scalable Research and Data Infrastructure for Open Science

VYACHESLAV TYKHONOV, Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Sciences
ANTON POLISHKO, CoronaWhy
ARTUR KIULIAN, CoronaWhy
MAKSYM KOMAR, CoronaWhy

At *CoronaWhy*[1] we are building a Common Research and Data Infrastructure for Open Science that can be used by researchers coming from various scientific communities involved in COVID-19 research. This distributed and scaled infrastructure follows *Reproducible Science*[5] and *FAIR*[10] principles and should be suitable for other important scientific challenges such as cancer and AIDS research. The vision of the community is to build this Artificial Intelligence infrastructure completely from Open Source components and with publicly available ML models like scispaCy[7] developed by Ai2 and other organizations. All data should be published and curated in *Dataverse*[2] where the provenance information is also available for every dataset. To make the pipeline reliable and verified by human experts, we are running two different annotations services, *Hypothesis*[3] for the evaluation of the statements extracted from *COVID-19* related papers and *Doccano*[4][6] for Natural Language Processing annotations.

The main challenge of this work is to get credibility and trust from all involved communities, especially from the medical experts as usually they don't have confidence in the results produced by people from other communities like Computer Science or Scientometrics. This research infrastructure effort should increase the involvement of the medical community in the analysis of *COVID-19* research papers and datasets, the transparency of data and services can guarantee the reproducibility of all experiments. *CoronaWhy* community is using Harvard Data Commons[4][1] as a foundation for all members to work together on the same problem and organizes efficient communication and collaboration through data exchange and reuse.

The final goal of CoronaWhy is to build and standardize Data Lake and interlink all independent COVID-19 Knowledge Graphs produced by different scientific communities sharing the same data, pipelines and services. Biological Knowledge Graph being developed by using *Biological Expression Language*[5] (BEL[3]) and helping researchers to find answers on COVID-19 related questions by using protein-protein interactions, protein functions and disease phenotypes. We are using the *Integrated Network and Dynamical Reasoning Assembler*(INDRA[2]) to assemble information about causal mechanisms and create predictive and explanatory models. With building of Social and Economic Knowledge Graphs it should be possible to understand the economic impact of coronavirus and investigate how quarantine and social distancing measures affected the population. The access to all available knowledge graphs will be delivered by appropriate CoronaWhy services like *Virtuoso*[6] with *SPARQL*[7] and *GraphQL*[8] endpoints and exposed as *Linked Data*[8].

This horizontal organization of the research infrastructure allows to form new vertical teams working on the specific research topic. The team members have a shared and collaborative access to all CoronaWhy data and tools through *JupyterLab Python/R notebook*[9] service and can work in the same workspace both individually or as part of own team. This open collaboration facilitating the high scale analysis of COVID-19 related entities and allows to reuse existent research tools and *dashboards*[9] developed, for example, for *data visualization*[11].

This bottom-up process of building CoronaWhy infrastructure can be considered as a lesson for other research infrastructures dealing with coronavirus data both in Europe and worldwide.

REFERENCES

[1] MERCÈ CROSAS. 2020. Harvard Data Commons. https://scholar.harvard.edu/mercecrosas/presentations/harvard-data-commons
[2] Benjamin M Gyori, John A Bachman, Kartik Subramanian, Jeremy L Muhlich, Lucian Galescu, and Peter K Sorger. 2017. From word models to executable models of signaling networks using automated assembly. *Molecular Systems Biology* 13, 11 (2017), 954. https://doi.org/10.15252/msb.20177651 arXiv:https://www.embopress.org/doi/pdf/10.15252/msb.20177651
[3] Charley Hoyt, Natalie Catlett, and Anselmo Di Fabio. [n.d.]. BEL: Biological Expression Language tools and services. https://bel.bio Software available from https://bel.bio.
[4] Gary King. 2007. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. *Sociological Methods and Research* 36 (2007), 173–199.
[5] M. Munafò, B. Nosek, and D. Bishop. 2017. A manifesto for reproducible science. *Nature Human Behavior* 1, 0021 (2017). https://doi.org/10.1038/s41562-016-0021
[6] Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. *doccano*: Text Annotation Tool for Human. https://github.com/doccano/doccano Software available from https://github.com/doccano/doccano.
[7] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics, Florence, Italy, 319–327. https://doi.org/10.18653/v1/W19-5034
[8] A.M. Scharnhorst, Marieke van Erp, Ronald Siebes, Christophe Dominique Marie Guéret, Tom Crick, Vyacheslav Tykhonov, Gerard Coen, Richard P. Smiraglia, P.K. Doorn, H. van den Berg, Jerry de Vries, A. Merono-Penuela, A. Ashkpour, and Reinier De Valk. 2019. *Curating and Archiving Linked Data Datasets from the Humanities - From Data of the Present to Data of the Future*. ADHO.
[9] Vyacheslav Tykhonov, Richard Zijdeman, and Jerry de Vries. 2015. Stakingen in kaart: Mapping strikes.
[10] Mark D. Wilkinson, Michel Dumontier, Susanna-Asunta Sansone, Luiz Olavo, Mario Prieto, Dominique Batista, Peter McQuilton, Tobias Kuhn, Philippe Rocca-Serra, Merc Crosas, and Erik Shultes. 2019. Evaluating FAIR maturity through a scalable, automated, community-governed framework. *Nature-Springer Scientific Data* 6, 174 (2019). https://www.nature.com/articles/s41597-019-0184-5
[11] Kalliopi Zervanou, Vyacheslav Tykhonov, A. van den Bosch, and Marien van der Heijden. 2014. Visualisation of 700 Years of Labour Conflicts in the Netherlands. 1–10.

[1] https://coronawhy.org
[2] https://datasets.coronawhy.org
[3] http://hypothesis.labs.coronawhy.org
[4] http://doccano.labs.coronawhy.org
[5] https://bel.labs.coronawhy.org
[6] https://virtuoso.openlinksw.com
[7] https://sparql.labs.coronawhy.org/sparql
[8] https://graphql.org
[9] http://colab.coronawhy.org

Authors' addresses: Vyacheslav Tykhonov, vyacheslav.tykhonov@dans.knaw.nl, Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Sciences; Anton Polishko, anton.polishko@gmail.com, CoronaWhy; Artur Kiulian, akiulian@gmail.com, CoronaWhy; Maksym Komar, komar@evologics.de, CoronaWhy.