

Computational ecosystems for social science*

Markus Kainu ^{†1}, Joona Lehtomäki ^{‡2}, Juuso Parkkinen ^{§3}, Juha Yrjölä ^{¶4}, Måns Magnusson ^{||5},
and Leo Lahti ^{**6}

¹Aleksanteri Institute, University of Helsinki, Finland

²Department of Biosciences, University of Helsinki, Finland

³Reaktor Innovations Oy, Finland

⁴Kansan Muisti ry, Finland

⁵Linköping University, Sweden

⁶Department of Veterinary Bioscience, University of Helsinki, Finland

December 7, 2014

Keywords: social science; elections; open government data; statistical programming; machine learning

Open availability of scientific and governmental data is increasing rapidly, creating novel opportunities for quantitative social sciences, data journalism, and citizen science. High-quality machine readable data sources relevant to social sciences are now becoming available across the globe related to political decision making, welfare, stratification, traffic, and other fields. In addition to social network analysis that has gained popularity in the field (Lazer et al., 2006), there are ample opportunities to expand the scope of computational social science research based on open governmental data sources. However, these resources are currently highly scattered and come in various formats, hindering their wider adoption. Various web-based tools for specific analysis tasks are available, but more flexible computational tools are urgently needed for more flexible and precise data processing and analyses. In particular, flexible algorithms are needed to clean up raw data and carry out fast expert-driven interactive data exploration and visualization.

There have been few attempts (King et al. 2013, Jockers 2013) within the academia to systematically bring the new data streams and computational methods into use in social science. We emphasize the importance of domain specific computational ecosystems in this task, and propose as one solution the rOpenGov community-driven ecosystem. It is based on the R statistical programming language, which has a strong track record in other fields, a versatile computational ecosystem and an active developer community. Within the R ecosystem there is a well established software release system, support for rich statistical simulation and modeling functions, and state-of-the-art visualization capabilities for addressing the diversity of analysis tasks in social sciences. With rOpenGov, we complement this existing ecosystem by providing a versatile set of R-based tools to access, preprocess and analyze open governmental data streams specifically relevant to social sciences.

The rOpenGov project builds on lessons learned from similar initiatives in other fields such as Bioconductor and rOpenSci. As these examples have shown, a community-driven open source approach is central as data availability is increasing, helping to pool scarce research resources. Using transparent research tools improve research reproducibility by providing standard workflows that can be easily adapted to different analytical tasks. The shared software ecosystem enables rapid development of extensible, scalable, and interoperable software, improvements through bug fixing, and provides tools to explore and expand the quantitative social science research methods.

The rOpenGov project launched in 2013 and is maintained by a core team of PhD-level computational scientists. The core team provides support for package authors and the user community by maintaining the infrastructure, reviewing new packages, and proposing recommended guidelines. An international contributor base has emerged from countries including the USA, Finland, Sweden, Austria, and Poland. The project is maintained in GitHub.

Election data analytics is a specific example of an active community project with implications for research and public outreach. We have created a public website kansanmuisti.fi/ to collect and organize openly available election related data in Finland, including a million social media feeds from election candidates collected since 2012, historical election

*Extended abstract submitted for International Conference on Computational Social Science in June 8-11, 2015 in Helsinki, Finland

[†]markuskainu@gmail.com

[‡]joona.lehtomaki@helsinki.fi

[§]juuso.parkkinen@iki.fi

[¶]juha.yrjola@iki.fi

^{||}mans.magnusson@gmail.com

^{**}leo.lahti@iki.fi

