Open Computational Ecosystems for Social Science

Contact: ropengov.github.io

Markus Kainu, Joona Lehtomäki, Juuso Parkkinen, Juha Yrjölä, Måns Magnusson, Mikko Tolonen, Niko Ilomäki, Leo Lahti

Open data creates opportunities for social sciences

The recent explosion in open data availability has created novel opportunities for research, journalism and citizenscience.
Taking advantage of these new data streams in computational social science, digital humanities, and related fields needs novel analytical tools. First software libraries are now emerging and have a huge potential to contribute to transforming these fields.

Power of open source communities

Efficient data analysis relies on customized, reproducible analysis workflows that are best developed jointly by the user community. Availability of ready-made algorithms for standard data analysis tasks allows an individual researcher to avoid reinventing the wheel, leaving more time to solve the specific research problems. Successful examples of open source communities have emerged in data intensive research fields, such as bioinformatics and particle physics, based on open source statistical programming languages.

Advantages of the open development model

- **Open source**: We use GitHub for shared version control. Openly licensed contributions guarantee that the tools are freely available for the international scientific community.
- **Reproducible documentation**: We provide online high-quality tutorials with fully reproducible documentation on how to access and analyse specific data sources, and to report the statistical results.
- **Transparent research**: The data analysis steps from raw data to the final results are published in full detail. To exemplify this, we publish reproducible case studies based on open data and algorithms in the rOpenGov blog.
- **Standardization**: A community-driven joint development approach helps to pool scarce research resources, develop common standards and ensure compatibilty for data analysis.

rOpenGov - An emerging ecosystem

rOpenGov is a statistical ecosystem focused on open source data analysis algorithms relevant to computational social sciences and digital humanities. We build on experiences learned from similar initiatives in other fields, such as Bioconductor and rOpenSci. We use the R statistical programming language, which has a versatile computational ecosystem with rich statistical modeling and state-of-the-art visualization capabilities.

Reproducible research blog

The rOpenGov blog at ropengov.github.io is an emerging collection of example case studies that showcase the opportunities of reproducible open data analytics. The general-purpose research algorithms with a wider applicability are distributed as open source R packages. This provides **well-documented tools to download, preprocess, integrate, analyse, visualize and report digital data streams in a fully automated and transparent fashion**. This complements the existing R ecosystem by **focusing on methodologies relevant to Computational Social Science and Digital Humanities**.

rOpenGov packages

Open Government Data

- Regions: eurostat
- Countries: Russia, Finland, Poland, USA
- Cities: Helsinki
- Statistics Authorities (Finland, Denmark, Sweden, PX-Web)
- GIS tools (Finland, OpenStreetMap, WFS..)
- Data anonymization
- Meteorology
- Health and Demography
- Political Science

Digital humanities and media

- Bibliographic analysis (Finland, UK, Europe)
- Media (Enigma, ProPublica, Sunlight Foundation, New York Times)

Parliamentary monitoring

- Election data (Austria, Finland, Russia, Huffpost Pollster)
- Quality of Government Institute

We are thankful for a number of developers for supporting this community. For a full list, see ropengov.github.io.

References

1. rOpenGov core team (2013). R ecosystem for open government data and computational social science. NIPS Machine Learning Open Source Software workshop (MLOSS). December 2013, Lake Tahoe, Nevada, US
2. S. Fortunato and C. Castellano (2012). Physics peeks into the ballot box. Physics Today 65:74
3. G. King, J. Pan and M. E. Roberts (2013). How Censorship in China Allows Government Criticism but Silences Collective Expression. American Political Science Review, 107(02), 326–343
4. M. L. Jockers (2013). Macroanalysis: Digital Methods and Literary History. University of Illinois Press.
5. S. Chou, W. Li and R. Sridharan, Democratizing Data Science.
6. D. Lazer, et al. (2009). Computational Social Science 323, 721–723

Eurostat open data: a reproducible example

```
plot(seq(6))
```