Open Data Science

Leo Lahti[1][0000−0001−5537−637X]

University of Turku, Finland
leo.lahti@iki.fi
http://www.iki.fi/Leo.Lahti

Abstract. The increasing openness of data, methods, and collaboration networks has created new opportunities for research, citizen science, and industry. Whereas openly licensed scientific, governmental, and institutional data sets can now be accessed through programmatic interfaces, compressed archives, and downloadable spreadsheets, realizing the full potential of open data streams depends critically on the availability of targeted data analytical methods, and on user communities that can derive value from these digital resources. Interoperable software libraries have become a central element in modern statistical data analysis, bridging the gap between theory and practice, while open developer communities have emerged as a powerful driver of research software development. Drawing insights from a decade of community engagement, I propose the concept of *open data science*, which refers to the new forms of research and research quality enabled and facilitated by open data, methods, and collaboration.

Keywords: Algorithmic Data Analysis · Open Data Science · Open Collaboration · Open Research Software

1 Introduction

Academic research is fundamentally a collective effort. The increasing openness of data, methods, and collaboration networks has created new opportunities to support and advance research, and this has been greatly facilitated by open developer communities. Virtual, and often informal research networks have emerged as powerful drivers of open research, and led to the formation of more structured communities and collaboration platforms such as the rOpenSci[1] and Bioconductor [5] that are expanding the scope of these efforts towards increasingly domain-specific algorithms.

Research software plays an essential role in bridging the gap between data, theoretical models and application expertise. Access to open data on various areas ranging from biomedical measurements and geospatial information to demographics, government activities, and historical records has opened new opportunities for research but there is a persisting shortage of algorithmic tools to

[1] https://ropensci.org

access, process and analyse open data resources. Domain-specific tools are often missing, and researchers often spend remarkable efforts on building custom scripts that never become widely distributed and verified despite their broader research potential. Developer communities can often provide social and technical incentives to promote long-term development and maintenance of open research software and collaboration networks, when immediate academic incentives are lacking.

Thriving virtual research communities have formed, for instance, around statistical programming environments, such as R, Python, or Julia, which have now rapidly growing ecosystems of research software and algorithms. Such open source ecosystems provide the means to standardize many routine tasks in data analysis, and building blocks for more comprehensive data science workflows. The open development model has emerged as a predominant mode for research algorithm development in natural sciences during the past decade, and is now rapidly gaining ground in the social sciences and humanities. The research potential of collective initiatives exceeds far beyond the capacities of any single research group or institution.

The field of *open data science*, proposed in this perspective article, refers to the new forms of research and research quality enabled by open data, algorithms, and collaboration networks in data-intensive fields. This includes both novel research opportunities that entirely based on open digital resources, as well as research projects that blend data and algorithms from both open and so far closed sources. Hence, open data science is a field that is both taking advantage of and advancing the development towards more open and collaborative research, which can support efficient, transparent and reproducible research through standardization and collective verification of common analysis tasks.

2 Elements of open data science

The three pillars of open data science include open data, open algorithms, and open collaboration (Fig. 1). Research software has a central role in mediating the interaction between these elements, helping to simplify, standardize, and automate research. Open collaboration networks can play a key role by providing peer support, collective quality control, and collaborative development opportunities [2, 4]. Techniques from the machine learning and artificial intelligence can support various steps of algorithmic data analysis from raw data access through analysis to final reporting [15]. Open data science emphasizes data and methods sharing through open infrastructures and virtual collaboration networks that have been enabled by digitalization and the push towards openness in government and academia.

2.1 Open data

A number of data repositories have been opened by academic, governmental, and industrial parties, and have been integrated with data analytical workflows

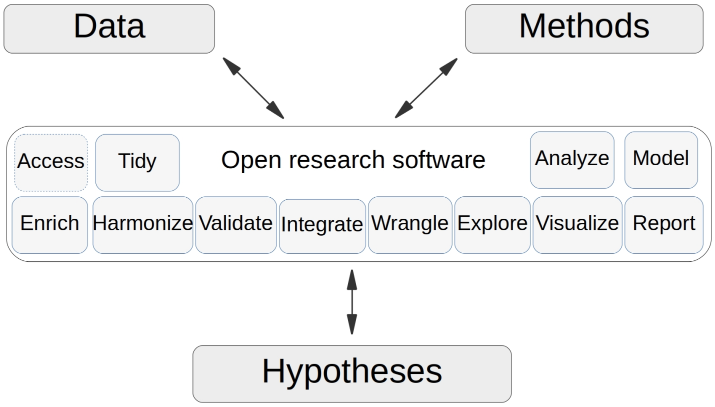

Fig. 1: Role of research software in the open data science workflow. Open research software mediates the community efforts and interaction based on shared data, models, and applications at each stage of an open data science project. Open collaboration can facilitate standardization and reproducibility, and thus the overall quality of research.

in research and commercial applications. Eurostat, the statistical office of the European Union, is one example of the many institutions that are sharing vast open data resources online[2]. The Eurostat database contains thousands of contemporary and longitudinal data sets on demography, economics, health, infrastructure, traffic and other topics at the European level, often with fine spatial and temporal resolution spanning several years or decades. Open data sharing differs from more traditional data services that provide access only to limited subsets of the data, and limit data analysis options on tools that are readily implemented in the query interface rather than letting the researcher decide which tools and analyses to execute. Such limitations form severe bottlenecks for research that relies on access to the full raw data.

The research use of open data, on the other hand, has been limited by the shortage of efficient tools to access, process, and integrate such data sets. The data sources and formats are scattered, and the methods for handling such data are heterogeneous, requiring a multitude of expertise and skills due to variability in data formats and interfaces. Research communities can benefit from shared programmatic tools that can seamlessly integrate initial data retrieval with downstream algorithms for statistical analysis and reporting. Standardized software libraries have been developed to facilitate fluent retrieval of open data from within statistical programming environments such as R, Python, or Julia. Collaborative development and automation of the open data retrieval is a central element in open data science.

[2] http://ec.europa.eu/eurostat/data/database

For instance, the *eurostat* R package [6] is specifically tailored to retrieve data in the R environment from Eurostat open data portal. The package includes custom tools to query, download, manipulate, and visualize these data sets in a smooth, automated and reproducible manner. Standard features, such as compliance with tidy data principles [18], support the integration with other tools of open data science. Significant portions of the package documentation have been published as open and reproducible case studies based on the Eurostat open data, providing concrete examples of possible research use, and a straightforward starting point for further adjustments. In the case of eurostat, also many generic database packages, such as *datamart* [17], *quandl* [12], *pdfetch* [13], and *rsdmx* [1], could be used to retrieve the open data sets. However, these more generic database packages are not dedicated to Eurostat data access, and do not therefore fully support the full spectrum of Eurostat open data services and their research use. This highlights the need to complement generic database tools with targeted algorithms that are specifically tailored to access particular data sources, and facilitate their integration with other data sets and statistical tools (Fig. 2).

Community projects, such as rOpenSci[3] and our own project, the rOpenGov [10][4], have emerged to provide various algorithmic tools and software packages for accessing open data portals. Many such packages are now distributed by open data science initiatives such as Bioconductor, CRAN, and rOpenSci, for instance. Such tools are typically created by the user communities, and facilitate data access independently of the original data provider.

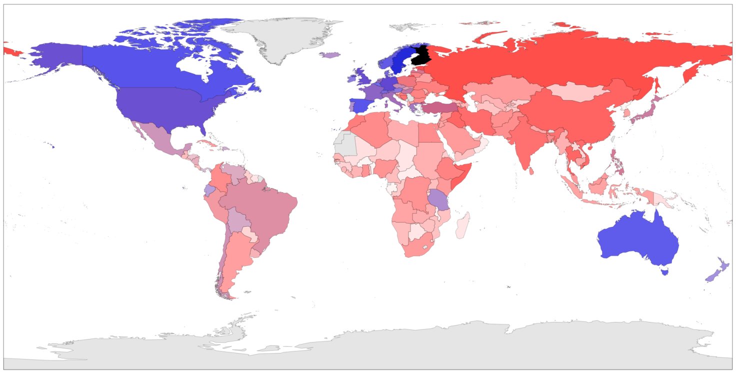

Fig. 2: Migration to (red) and from (blue) Finland in 2011 according to data from Statistics Finland as retrieved with the *pxweb* R package. The visualization relies on custom tools in the R statistical programming environment, from raw data access to harmonization, statistical analysis, and visualization.

[3] https://ropensci.org

[4] http://ropengov.github.io

2.2 Open algorithms

The increased open data availability has potential to support and renew research but the methodological basis of open data science is still shaping up. Tools for standardized open data retrieval need to complemented by algorithms for statistical data exploration, analysis, and modeling (Fig. 3). Statistical programming environments provide access to a vast body of advanced statistical techniques for data analysis, including techniques such as (generalized) linear models, statistical machine learning methods, probabilistic programming [3, 14], data integration, and visualization techniques [19].

Many research projects rely on rich combinations of spatio-temporal, textual, personal, demographic, and other types of data that may require remarkable amounts of dedicated custom processing before systematic and reliable statistical analysis becomes possible. Hence, general-purpose methods need to be complemented by data processing and analysis algorithms that support research on particular research areas. Most methods for advanced statistical analysis and modeling assume that data is readily available in a clean or tidy format. This does not hold in most real research situations. Hence, data cleaning and harmonization often forms a major component in research projects. Projects such as the *tidyverse* [20] have emerged to harmonize and organize research data before and during various stages of statistical analysis. Such general-purpose data wrangling methods can be complemented by domain-specific tools for customized data subsetting and manipulation (see e.g. [11]), however. Our experience is in line with the frequently encountered statement that the majority of the effort in data science projects is spent on organizing and harmonizing data before it is amenable for research use. In practice, data cleaning and harmonization often rely on combinations of automation and manual work. Intelligent algorithms for data analysis can greatly benefit from domain-specific tools for data wrangling, subsetting and visualization.

Our recent work on the historical development of print press provides an example [7], where we developed algorithmic tools to clean up bibliographic metadata collections in a scalable manner. This is now allowing a detailed quantitative analysis of historical book production across Europe. In order to estimate paper consumption, for instance, we extracted information on books heights, widths, and page counts. This included converting various standard book formats such as folio, quarto etc. into the SI system, summing up information on cover pages, special pages, and so forth. Furthemore, we augmented the data and analyses with publicly available information on name-gender mappings, author metadata, and other sources of public information. The open algorithms can be verified and further improved when potential inconsistencies are observed, and the data sets can be gradually refined over time when new information arrives. Replacing manual curation by supervised machine learning techniques is now helping to scale up this research to cover millions of print products. We anticipate that the demand for such open and customized analysis methods and workflows will increase rapidly in this field when research libraries start to share these data resources more openly [7].

Finally, data harmonization and statistical analysis need to be complemented with high-quality visualization and reporting. Published visualizations often rely on geospatial maps or demographic data that are available from multiple governmental and international institutions. The eurostat package, for instance, can be used to download custom administrative boundaries by EuroGeographics, thus supporting seamless data visualization on the European map. Similar geospatial tools are also available for specific countries and cities, for instance[5].

Collaboratively developed tools to access, harmonize, integrate, and analyse large data collections are needed to pool scarce resources and increase the efficiency of data-intensive research. Open collaboration networks can gradually accumulate and refine collections of targeted algorithms in open statistical programming environments, as demonstrated by Bioconductor, rOpenSci, and other existing open data science projects. Hence, intelligent data analysis is often critically dependent on the overall data analytical infrastructures that provide the fundamental context for the application of state-of-the-art analysis algorithms.

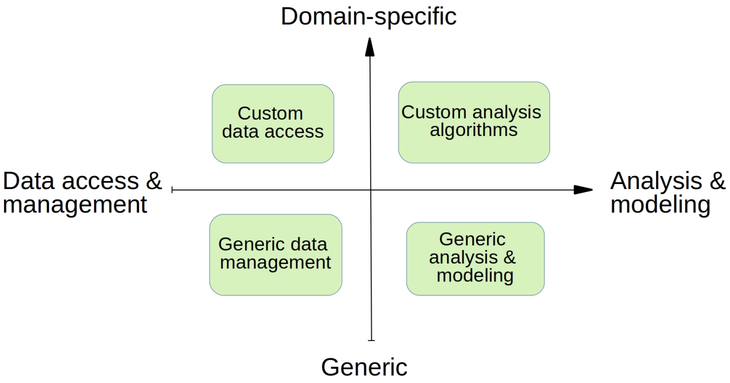

Fig. 3: Tradeoffs in open data science. The emphasis in methods development is shifting from standard data management towards intelligent algorithms for data analysis and modeling (horizontal axis). At the same, general-purpose tools are increasingly complemented by targeted domain-specific algorithms (vertical axis). Open data science aims at standardization but this is contrasted with the constant tendency to drift towards vast flexibility in the development of alternative methods and innovative combinations of the shared data and software components. Increasing openness of data and algorithm is bringing up new opportunities for research and collaborative methods development.

[5] The gisfin and helsinki packages; see http://ropengov.github.io

2.3 Open collaboration

Whereas open data science emphasizes the role of collaborative methods development and refinement [2, 9], the efforts to standardize data analytical methods are balanced by a constant drift towards flexibility and custom methods (Fig. **??**). Open and institutional community projects have emerged to balance these needs based on R, Python, and other programming languages, resulting in vast networks of developers and users of open research software from natural sciences [5] to humanities (e.g. rOpenSci). Related examples include the The OpenML [16] that provides tools to bring together data, algorithms and analysis results for open evaluation, and Project Open Data[6], which promotes the development and use of tools on open government data. We have made many such tools available within the rOpenGov project for computational social science and digital humanities. This is an example of a community-driven open data analytical ecosystem, which is now facilitating research use of many open institutional data resources based on a collection of over 20 R packages in varying stages of development. The eurostat package, for instance, evolved gradually from the earlier work by the same authors. Over time, multiple contributors joined in, and the package was extensively being developed and tested by various users before its eventual release. Open developer communities can also organize software review, promote data and software citation best practices, develop improved methods for authorship determination, and gain additional visibility for the projects. Academic teaching can also benefit from well-documented and reproducible workflows. In our experience, application specialists with little programming experience have been able to adopt practical skills, key tools and best practices using reproducible notebooks as an interactive learning tool.

3 Conclusion

This brief perspective introduced the concept of open data science. This new research methodology is emerging at the intersection of open data, methods, and collaboration networks. Open data science has become increasingly central for the overall quality and efficiency of data-intensive research, helping to renew and complement research in data-intensive fields. Open research practices are now transforming the way we understand and share research outputs [8, 9]. Open developer communities can provide social and technical incentives to promote open, collaborative work when academic incentives are lacking. Social aspects remain among the greatest challenges towards further development of open data science, as balancing the collaborative need for long-term development and maintenance with the prevailing authorship and incentive structures in academia remains a constant challenge. Given the enormous significance of high-quality open source software in modern data-intensive research, academic institutions and funding bodies should continue to develop and experiment with new ways to support sustainable development of open data science for instance

[6] https://project-open-data.cio.gov

by providing funding and recognition for the developers, maintainers and contributors of open research software, which is a key mediator of open data science. The use of statistical programming facilitates open participation, and allows full flexibility in constructing custom workflows in order to harness the full potential of modern data analysis and visualization arsenal. These methods can be used and further tested and refined by the research community as well as other parties, contributing to the growing open source ecosystems in natural sciences, social sciences, and digital humanities. Emphasis on open research practices will help the research communities to avoid replication to pool scarce research resources, and find improved methods for collective analysis and verification of research hypotheses.

Acknowledgements

I am grateful to the rOpenGov contributors, in particular Joona Lehtomäki, Markus Kainu, and Juuso Parkkinen, and our close collaborator Mikko Tolonen. The work has been partially funded by Academy of Finland (decisions 295741, 307127).

References

1. Blondel, E.: rsdmx: Tools for Reading SDMX Data and Metadata (2018). doi: 10.5281/zenodo.1173229, R package
2. Boettiger, C., Chamberlain, S., Hart, E., Ram, K.: Building software, building community: lessons from the rOpenSci project. Journal of Open Research Software **3** (2015). doi: 10.5334/jors.bu
3. Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., , Riddell, A.: Stan: A probabilistic programming language. Journal of Statistical Software **76** (2017). doi: 10.18637/jss.v076.i01
4. Gandrud, C.: Reproducible research with R and R Studio. Chapman & Hall/CRC (2013)
5. Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K.D., Irizarry, R.A., Lawrence, M., Love, M.I., MacDonald, J., Obenchain, V., Oleś, A.K., Pagès, H., Reyes, A., Shannon, P., Smyth, G.K., Tenenbaum, D., Waldron, L., Morgan, M.: Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Meth. **12**, 115–121 (2015). doi: 10.1038/nmeth.3252
6. Lahti, L., Huovari, J., Kainu, M., Biecek, P.: Retrieval and analysis of eurostat open data with the eurostat package. The R Journal **9**, 385–392 (2017), https://journal.r-project.org/archive/2017/RJ-2017-019/index.html
7. Lahti, L., Ilomäki, N., Tolonen, M.: A quantitative study of history in the English short-title catalogue (ESTC) 1470-1800. LIBER Quarterly **25**, 87–116 (12 2015). doi: 10.18352/lq.10112
8. Lahti, L., da Silva, F., Laine, M.P., Lhteenoja, V., Tolonen, M.: Alchemy & algorithms: perspectives on the philosophy and history of open science. RIO Journal **3**, e13593 (2017). doi: 10.3897/rio.3.e13593

9. Laine, H., Lahti, L., Lehto, A., Ollila, S., Miettinen, M.: Beyond open access - the changing culture of producing and disseminating scientific knowledge. In: Proceedings of the 19th International Academic Mindtrek Conference in Tampere, Finland, September 22-24. AcademicMindTrek'15: Proceedings of the 19th International Academic Mindtrek Conference, ACM, ACM New York, NY, USA (2015), http://dl.acm.org/citation.cfm?id=2818187
10. Leo Lahti, Juuso Parkkinen, J.L., Kainu, M.: rOpenGov: open source ecosystem for computational social sciences and digital humanities (2013), http://ropengov.github.io, ICML/MLOSS workshop (Int'l Conf. on Machine Learning - Open Source Software workshop).
11. McMurdie, J., Holmes, S.: phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE **8**, e61217 (2013). doi: 10.1371/journal.pone.0061217
12. McTaggart, R., Daroczi, G., Leung, C.: Quandl: API wrapper for quandl.com (2015), http://CRAN.R-project.org/package=Quandl, R package version 2.7.0
13. Reinhart, A.: pdfetch: Fetch Economic and Financial Time Series Data from Public Sources (2015), http://CRAN.R-project.org/package=pdfetch, R package version 0.1.7
14. Salvatier, J., Wiecki, T., Fonnesbeck, C.: Probabilistic programming in python using pymc3. PeerJ Computer Science **2**, e55 (2016). doi: 10.7717/peerj-cs.55
15. Toivonen, H., Gross, O.: Data mining and machine learning in computational creativity. Wiley Int. Rev. Data Min. and Knowl. Disc. **5**, 265–275 (2015). doi: 10.1002/widm.1170
16. Vanschoren, J., van Rijn, J.N., Bischl, B., Torgo, L.: OpenML: networked science in machine learning. SIGKDD Explor. Newsl. **15**, 49–60 (2014)
17. Weinert, K.: datamart: Unified Access to your Data Sources (2014), http://CRAN.R-project.org/package=datamart, R package version 0.5.2
18. Wickham, H.: Tidy data. Journal of Statistical Software **59** (2014). doi: 10.18637/jss.v059.i10
19. Wickham, H.: ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York (2016), http://ggplot2.org
20. Wickham, H.: tidyverse: Easily Install and Load the 'Tidyverse' (2017), https://CRAN.R-project.org/package=tidyverse, R package