A practical guide to developing research objects when undertaking reproducible statistically orientated social science research during COVID-19

The unprecedented nature of the COVID-19 global pandemic has had extremely disruptive effects on contemporary social life. The empirical findings that flow from social science inquiries have important implications for establishing policies and changing practices. The speed at which the pandemic has unfolded has led to a previously unparalleled requirement for rapid results from social science studies. This acceleration has consequences for verifying empirical results, and for building incrementally on research findings.

In another document we provide general guidance on how to adopt transparent and reproducible practices in statistically orientated social science research during the COVID-19 pandemic (http://eprints.ncrm.ac.uk/4402/). One recommendation was the production of research objects. Research objects are uncommon in the social sciences and they are introduced and explicated in this guide.

Background

The COVID-19 pandemic is an urgent threat to global health. During the pandemic a number of authors in the scientific community have emphasised that research must be a reliable, rigorous and transparent process, because research findings need to be rapidly translated into practices[1].

In health research it has been recognised that when researchers share data, research code, and software, and generally make their work as transparent as possible, it allows other researchers to verify results and to expand upon work and public officials to make scientifically informed decisions[2]. Similarly, social science research on COVID-19 must be transparent in order that findings can be verified, and that results can be reproduced and incrementally developed[3].

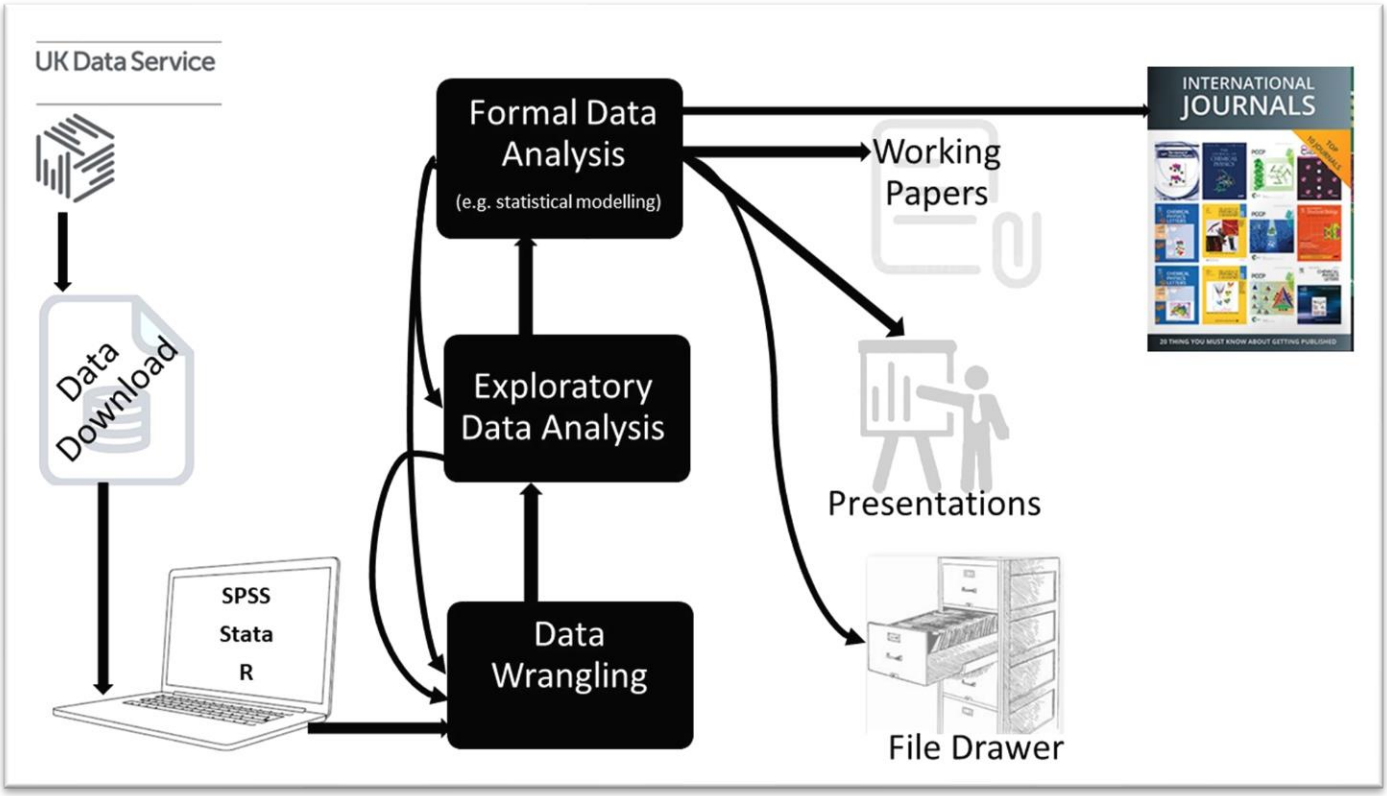
The NCRM have provided general guidance on how to adopt transparent and reproducible practices in statistically orientated social science research during COVID-19 (http://eprints.ncrm.ac.uk/4402/). One recommendation is the production of research objects. Currently, this practice is little known in social science, and it is explained in this guide. This guide was produced during the COVID-19 pandemic, and reflects current methodological thinking.

The Workflow in Statistically Orientated Social Science Research

The term ‘workflow’ describes the co-ordinated framework of activities required for conducting statistically orientated social science research[4]. The workflow spans the entire process from planning the research through to publishing research outputs. The workflow typically includes acquiring the data, wrangling the data, exploratory data analysis, formal data analyses (e.g. statistical modelling), and the production of both informal outputs (e.g. presentations) and formal outputs (e.g. working papers and journal articles) (see figure 1).

It is implausible for social researchers to expect to undertake any serious statistically orientated research without using a computer and a data analysis software package, or a statistical programming language. Software can be operated in different ways. Graphical user interfaces (such as drop down menus) do not provide a suitable record of the very large number of actions in the data enabling process (often called data wrangling) that are required to produce the ‘analytical

Figure 1 The Typical Statistically Orientated Social Science Workflow



dataset' that is required for the empirical research. In practice researchers must write out the code for software commands using a programming or syntactical format[5].

Central to a transparent and reproducible workflow is a clear audit trail. The goal of the audit trail is to document the provenance of every result [4, 6]. For many researchers the backbone of audit trail will be their files of programming commands or syntax (e.g. a .do file in Stata, a .R file containing R code, or a .sav file in SPSS). A transparent public audit trail allows researchers who are unconnected with the original research to gain access to all of these stages of the research process. This is imperative for COVID-19 social research because it enables the rapid verification of findings and for results to be swiftly build upon.

Research Objects

Within data science the concept of a research object (RO) describes an artefact that packages up research outputs (e.g. data, metadata, code, results, documentation, and papers) [7, 8].

Expressed formally, research objects are rich aggregations of resources, that possess some scientific intent or support a research objective[7]. In practice research objects can be considered as bundles of the digital bits and pieces that make up the reusable record of a piece of research, and they are identifiable, citable and sharable[9].

Research Objects for Statistically Orientated Social Science Research

A research object for a statistically orientated social science research output is likely to comprise the following.

1. A summary document that narrates the research object, for example a README file.

2. The analytical dataset in an open and readable format, if it is legal, ethical and feasible to include it within the research object. In practice, social scientists working with large-scale datasets, for example those supplied by national archives, will be unable to share

data. This is because the data are provided under some form of 'end user license' that prevents data sharing[1].

3. A link that clearly identifies the exact version of the unprocessed (or raw) dataset and its origins (i.e. where and when it was obtained) using a persistent identifier such as a digital object identifier (DOI)[10]. This acts as a substitute for the analytical dataset when it cannot be included as part of the research object.

4. A statement indicating which data analysis software package, or a statistical programming language was employed, which clearly states the version, and all the libraries, dependencies and plugins that were used. This should be accompanied by detailed information about the computer that was used and the computational environment in which the work was undertaken[11].

5. Files of programming commands or syntax that were used for acquiring the data, wrangling the data, exploratory and formal data analyses, and the production of publication outputs (e.g. graphs and tables). For example this will be a .do file in Stata, a .R file containing R code, or a .sav file in SPSS. The 'file drawer problem' is a term used to describe the detrimental consequences that the under-reporting of non-significant research findings has on the landscape of empirical research[12]. The 'file drawer problem' is partially addressed because the files in the research object will document all of the analyses undertaken, and not just the analyses that are presented in the published work. The files of programming commands may be organised into an electronic research notebook, for example a Jupyter notebook[13, 14].

6. A set of intermediate research outputs. For example, these may include slides from presentations and working papers.

7. Research outputs, for example academic journal articles. In practice this may be via gold open access or green open access (e.g. through a university repository).

Research Repositories

The Open Science Framework (OSF) is a specialist platform which provides collaboration tools that help researchers both to work on projects privately, and to make entire projects publically accessible for broad dissemination[15]. Research objects can easily be constructed on OSF because files of research code can be shared alongside further project related materials such as conference presentations and preprints. It is currently in infancy, but the OSF platform shows promising signs that it could emerge as a dominant eco-system for transparent and reproducible social science.

GitHub is one possible alternative to OSF. GitHub is primarily a software development platform[16]. The functionality of GitHub lends itself well to developing public repositories of social science workflows.

FAIR Principle for Social Research

The production of research objects should be guided by **FAIR** principles, this means that they should be **F**indable, **A**ccessible, **I**nteroperable and **R**eusable[17]. These principles should assist the discovery and reuse by third-parties that are unconnected with the original research[18].

A social science research object is **F**indable when it can be uniquely and persistently identified. Elements of a social science research object are **A**ccessible if they can be obtained by other researchers or stakeholders such as policy makers. The information contained in the research object must be easy to access. Files containing research code must be accessible to other computers.

Elements of a social science research object are **I**nteroperable when they are understandable and allow exchange. For example, a file in an esoteric format (e.g. .xzq) would not be understood by human researchers or readable by another computer. Therefore, the information within the file could not be exchanged.

Components of a social science research object are **R**eusable when they are sufficiently well described that they can be utilized by a third party unconnected with the original research with minimal effort.

[1] See https://www.ukdataservice.ac.uk/conditions.aspx accessed 22.03.20 for detailed information on data supplied by the UK Data Service.

Useful resources

https://www.researchobject.org/

https://www.force11.org/group/fairgroup/fairprinciples

Connelly, Roxanne, Vernon Gayle, and Chris Playford. Transparent and Reproducible Data Analysis. SAGE Publications Limited, 2020. https://methods.sagepub.com/foundations/transparent-and-reproducible-data-analysis

Playford, Christopher J., Vernon Gayle, Roxanne Connelly, and Alasdair JG Gray. "Administrative social science data: The challenge of reproducible research." Big Data & Society 3, no. 2 (2016): 2053951716684143. https://journals.sagepub.com/doi/pdf/10.1177/2053951716684143

Reproducible Social Research NCRM Online Resource by Vernon Gayle https://www.ncrm.ac.uk/resources/online/all/?id=20732

Research Object examples for 'Parental Social Class and Filial School Level Educational Outcomes in Contemporary Britain' ESRC SDAI PROJECT ES/R004978/1 https://osf.io/vgfnr/

References

1 Besançon, Lonni, Nathan Peiffer-Smadja, Corentin Segalas, Haiting Jiang, Paola Masuzzo, Cooper A Smout, Eric Billy, Maxime Deforet, and Clémence Leyrat. "Open Science Saves Lives: Lessons from the Covid-19 Pandemic." *BioRxiv* (2020).

2 Sumner, Josh Q, Leah Haynes, Sarah Nathan, Cynthia Hudson-Vitale, and Leslie D McIntosh. "Reproducibility and Reporting Practices in Covid-19 Preprint Manuscripts." *medRxiv* (2020).

3 Moon, M Jae. "Fighting Covid-19 with Agility, Transparency, and Participation: Wicked Policy Problems and New Governance Challenges." *Public administration review* 80, no. 4 (2020): 651-56.

4 Long, J.S. *The Workflow of Data Analysis Using Stata*. College Station: Stata Press, 2009.

5 Gayle, VJ, and PS Lambert. "The Workflow: A Practical Guide to Producing Accurate, Efficient, Transparent and Reproducible Social Survey Data Analysis." (2017).

6 Sandve, Geir Kjetil, Anton Nekrutenko, James Taylor, and Eivind Hovig. "Ten Simple Rules for Reproducible Computational Research." *PLoS computational biology* 9, no. 10 (2013): e1003285.

7 Bechhofer, Sean, Iain Buchan, David De Roure, Paolo Missier, John Ainsworth, Jiten Bhagat, Philip Couch*, et al.* "Why Linked Data Is Not Enough for Scientists." *Future Generation Computer Systems* 29, no. 2 (2013): 599-611.

8 Sefton, Peter, Eoghan Ó Carragáin, Carole Goble, and Stian Soiland-Reyes. "Introducing Ro-Crate: Research Object Data Packaging." In *eResearch Australia*. Brisbane, 2019.

9 De Roure, David. "Towards Computational Research Objects." Paper presented at the Proceedings of the 1st International Workshop on Digital Preservation of Research Methods and Artefacts, 2013.

10 Paskin, Norman. "Digital Object Identifier (Doi®) System." *Encyclopedia of library and information sciences* 3 (2010): 1586-92.

11 Howe, Bill. "Virtual Appliances, Cloud Computing, and Reproducible Research." *Computing in Science & Engineering* 14, no. 4 (2012): 36-41.

12 Salkind, Neil J. *Encyclopedia of Research Design*. Vol. 1: Sage, 2010.

13 Toomey, D. *Learning Jupyter*. Birmingham: Packt Publsihing, 2016.

14 Kluyver, Thomas, Benjamin Ragan-Kelley, Fernando Pérez, Brian E Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley*, et al.* "Jupyter Notebooks-a Publishing Format for Reproducible Computational Workflows." Paper presented at the ELPUB, 2016.

15 Foster, Erin D., and Ariel Deardorff. "Open Science Framework (Osf)." *Journal of the Medical Library Association: JMLA* 105, no. 2 (2017): 203.

16 Blischak, John D, Emily R Davenport, and Greg Wilson. "A Quick Introduction to Version Control with Git and Github." *PLoS computational biology* 12, no. 1 (2016).

17 Boeckhout, Martin, Gerhard A Zielhuis, and Annelien L Bredenoord. "The Fair Guiding Principles for Data Stewardship: Fair Enough?". *European journal of human genetics* 26, no. 7 (2018): 931-36.

18 Wilkinson, Mark D, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg*, et al.* "The Fair Guiding Principles for Scientific Data Management and Stewardship." *Scientific data* 3 (2016): 10.1038/sdata.2016.18.

This guide was produced in 2021 by Vernon Gayle, in collaboration with Roxanne Connelly and Christopher Playford and draws on work undertaken as part of ESRC Project ES/R004978/1 ‘Parental Social Class and Filial School Level Educational Outcomes in Contemporary Britain’ https://osf.io/vgfnr/.

National Centre for Research Methods
Social Sciences
University of Southampton
Southampton, SO17 1BJ
United Kingdom.

Web www.ncrm.ac.uk
Email info@ncrm.ac.uk
Tel +44 23 8059 4539
Twitter @NCRMUK