A Data Paper in the Social Sciences: Why, For Whom, and How?

Victor Gay[1]

Chapter prepared for *Publier, partager, réutiliser les données de la recherche: les data papers et leurs enjeux*, edited by Joachim Schopfel and Christine Kosmopoulos (Lille: Presses universitaires du Septentrion, in press). It is based on a presentation given at the #dhnord2021 conference in November 2021. Own translation.

Why write a data paper? While the humanities and social sciences have experienced a quantitative turn over the past decade, the scientific contribution of data production remain undervalued: promotion committees still favor traditional research articles and peers, while eager to reuse data produced by others, do not always credit them with proper citation. This lack of recognition seems incompatible with the time-consuming task of documenting the data production process and ensuring that the data produced conforms to the FAIR principles – both of which are necessary if research is to be reproducible. In this context, data papers are a tool that can help data producers gain recognition for their scientific contribution not only by making their data easier to cite, but also by improving the relevance and scope of the reuse of their data. Second, who should a data paper be written for? Currently, only a few journals in the humanities and social sciences welcome this editorial format. Paradoxically, this gap may represent an opportunity for data producers to reach a relatively broad and interdisciplinary audience. However, such broad reach requires adapting the writing to a non-specialist audience. It may also have implications for the choice of data format and repository. Finally, how to write a data paper such that it provides a comprehensive understanding of the data produced and proper guidance to their reuse? A solution is to draw from proven models from the hard sciences and adapt them to the specificities of the humanities and social sciences. This chapter highlights that a variety of skills are needed to produce a coherent "database-data paper" ecosystem. Fortunately, institutional support may help researchers acquire these skills and guide them throughout the process.

In this chapter, I draw on my experience in writing a data paper: "Mapping the Third Republic. A Geographic Information System of France (1870–1940)" (Gay 2021). This data paper describes a geographic information system of France during the Third Republic – the TRF-GIS database.[2] This database provides annual nomenclatures and shapefiles of the administrative constituencies of France from 1870 to 1940. It describes general administrative constituencies (*départements*, *arrondissements*, *cantons*) as well as military, judicial, penitentiary, electoral, academic, ecclesiastical, and labor inspection constituencies. It also provides annual nomenclatures that map each *commune* to the constituencies it belonged to.[3]

A Data Paper in the Social Sciences: Why?

The use of data has taken on a central role in humanities and social sciences research over the past decade, due in part to the unprecedented production of statistics on social facts and their availability through online data catalogs such as PROGEDO-ADISP.[4] This is for instance the case in sociology, a field that is at the heart of the social sciences. An analysis of the 400 articles published between 2000 and 2020 in the *Revue française de sociologie* reveals a clear trend towards quantitative

[1] Toulouse School of Economics (TSE) and Institute for Advanced Study in Toulouse (IAST), University Toulouse Capitole, Toulouse, France. Email: victor.gay@tse-fr.eu.

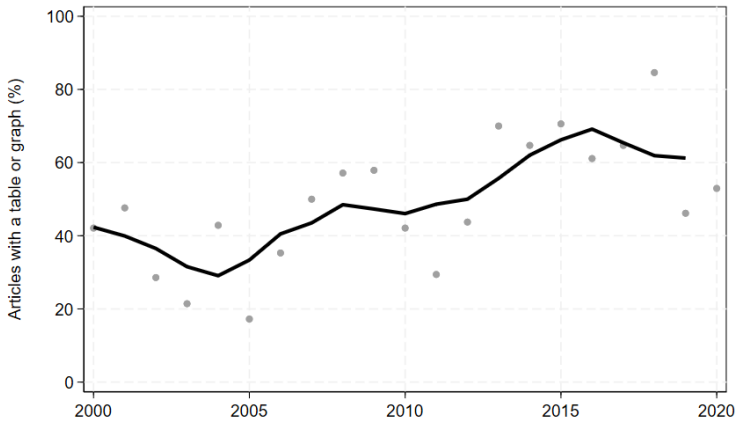
[2] TRF-GIS stands for *Third-Republic France Geographic Information System*.

[3] The data paper is freely available on HAL at https://hal.archives-ouvertes.fr/hal-02951461. The data are available on the Harvard Dataverse at https://dataverse.harvard.edu/dataverse/TRF-GIS.

[4] The PROGEDO-ADISP catalogue disseminates surveys and databases of French public statistics. It is accessible at http://www.progedo-adisp.fr/.

approaches: since a decade, more than half of articles in this journal contain at least one table or figure with statistics (Figure 1). This is also the case in the field of history, as highlighted by the issue of *Annales. Histoire, sciences sociales* devoted to quantitative history (Karila-Cohen *et al.* 2018), as well as the success of the book *Méthodes quantitatives pour l'historien* (Zalc and Lemercier 2008), recently republished in English under the title *Quantitative Methods in the Humanities* (Zalc and Lemercier 2019). This quantitative turn is also evident in Anglo-Saxon research, both in social and economic history: while the proportion of articles with at least one table or graph presenting statistics remained stable at around 90 percent in the main economic history journals between 2005 and 2020 (Cioni *et al.* 2021, p. 24), it rose from 5 to 13 percent in the *American History Review* over the same period (Ruggles 2021, p. 14).[5]

Figure 1. Proportion of articles in the *Revue française de sociologie* with at least one table or graph with statistics (2000–2020)



The curve represents a five-year moving average. Articles were collected via the Cairn.info portal for volumes 44 to 62 (2004–2021) and the Persée portal for volumes 39 to 43 (1998–2002). Editorials, notes, reviews, debates, translated articles, commentaries, and *in memoriam* articles are not included.

It is in this context that the reproducibility crisis – which began in psychology and medicine – caught up with the social sciences, starting with economics (Maniadis, Tufano, and List 2017). Because of various publication biases – *p*-hacking, low statistical power, confirmation biases on the part of authors and reviewers, etc. – most published research findings may in fact be "false positives," or in other words, a statistical illusion (Ioannidis 2005). Several responses have been proposed to alleviate this crisis: data management plans, pre-analysis plans, or even meta-analyses (Maniadis and Tufano 2017; Christensen and Miguel 2018). However, it seems that a necessary condition for overcoming the crisis is reproducibility *per se*, i.e., the ability to reproduce the results of published studies. This is not yet the case: for instance, Chang and Li (2022) show that only half of a sample of 67 macroeconomics articles published in leading journals are reproducible. The first step to solving the crisis is therefore to create the conditions for the availability and reuse of research data.

How can this be achieved? This is where the FAIR principles come in: reproducibility primarily requires data to be findable, accessible, interoperable, and reusable (Wilkinson *et al.* 2016). Abiding by these principles has been at the heart of French national research policy for several years, e.g., through the National Plan for Open Science and the funding of research by the ANR (CoSO 2019). However, regardless of the support provided by research institutions and

[5] The journals analyzed by Cioni *et al.* (2021) are the *Economic History Review*, the *Journal of Economic History*, *Explorations in Economic History*, the *European Review of Economic History*, and *Cliometrica*.

infrastructures, the burden of documenting and ensuring data compliance with FAIR principles ultimately falls on data producers themselves, i.e., researchers, who are already overwhelmed by administrative tasks that keep on accumulating (Ali and Rouch 2013). But this role is at odds with the incentives they face. Indeed, data production is generally not valued by promotion committees, which rely heavily on traditional publications in peer-reviewed journals (Gozlan 2016). Nor is data production valued by peers, who generally do not cite the data they reuse. For instance, Robinson-García *et al.* (2015, pp. 29–70) show that only 18 percent of the data reused in social science articles published between May and June 2013 and available on Web of Science were explicitly cited.[6]

In the face of these challenges, data papers offer interesting prospects. Indeed, by describing the data production process in an article published in a journal, they offer data re-users a simple means of citing the data, thereby providing scientific recognition to the work of data producers for promotion committees as well as peers – besides a reuse that is both more appropriate thanks to the documentation and more accessible thanks to compliance with the FAIR principles.[7] What's more, insofar as data papers are peer-reviewed, they generate incentives for data producers themselves to refine their data and descriptions, which in the end can only enhance data dissemination and the scope of their reuse (Walters 2020). In this sense, data papers can help solve the classic free-rider problem that characterizes the production of public goods – in this case, data that is properly documented and FAIR.

The data paper "Mapping the Third Republic" (Gay 2021) addresses these questions in the specific context of Third-Republic France (1870–1940). Indeed, this historical period was characterized by an unprecedented production of statistics (Desrosières 2010 [1993]) by administrations operating at heterogeneous levels of aggregation, in a context of profound socio-economic changes such as the Second Industrial Revolution (1870–1914) or the First World War (1914–1918). Thanks to new digitization techniques for statistical archives, quantitative historical research about this period is booming. However, the analysis and visualization of localized historical data requires a common reference framework or a Geographic Information System (GIS). In the absence of such a system for France during the Third Republic, each researcher had to tackle this task individually. This implied substantial loss of time, approximations due to the difficulty of the task, but also a lack of interoperability with other research programs and, ultimately, a lack of reproducibility. In fact, these information systems were rarely made available to the public and even more rarely complied with the FAIR principles.[8] The TRF-GIS database therefore offers a solution to these problems by providing FAIR data, which construction process is documented in a data paper.

A Data Paper in the Social Sciences: For Whom?

Who is a data paper written for? The potential users of the data described, of course. However, the readership of a data paper is often much broader and more interdisciplinary than the data producer originally anticipated. In fact, journals that accept this type of papers are still rare in the humanities and social sciences. There are currently two journals dedicated to data papers – data journals – in these fields: the *Journal of Open Humanities Data* and the *Research Data Journal for the Humanities and Social Sciences*—although some data journals publish articles in the social sciences and other

[6] A recent study on data citation practices in the field of biodiversity in 2019 shows that on a random sample of 100 articles reusing data, only 27 percent explicitly mention them in their reference lists (Khan *et al.* 2021).

[7] Even if the effect of data papers on data reuse have not reached their full potential yet, at least in the hard sciences (Jiao et Darch 2020).

[8] An exception are the canton shapefiles for 1884 and 1925 published by the LARHRA (2011) and constructed following the methodology of manual vectorization of historical georeferenced maps. See Gay (2021, pp. 14–15) for a critical analysis of this method.

disciplines, such as *Data in Brief* or *F1000 Research*.[9] In addition, some humanities and social sciences journals accept data papers in addition to traditional research articles, such as *Cybergeo* for geography, *Historical Methods: A Journal of Quantitative and Interdisciplinary History* for history, or *Frontiers in Sociology* for sociology. Finally, several journals offer the possibility of publishing data papers in their short articles section, without making an explicit distinction between this type of article and a traditional one. In history, this is for instance the case for *Explorations in Economic History* and *Histoire & mesure*.

The author of a data paper therefore has to face a limited number of outlets, which necessarily implies a wider readership than his or her own field of research. Such a scope, however, requires that the writing be adapted to a non-specialist audience: the context must be thoroughly presented so as to avoid disciplinary jargon, assumptions underlying the data collection and categorization must be made explicit, and the potential reuses of the data beyond one's own field of research must be clarified. These considerations also have implications for the format of the data: data can of course be disseminated in the format most commonly used in one's own discipline (SPSS, Stata, or SASS), but it is important to provide other fields the opportunity to use them by also proposing open and universal formats such as TXT or CSV. The choice of data repository must also be carefully considered: one can focus on a French repository such as Nakala or PROGEDO-ADISP, or aim for a more international audience by using the Harvard Dataverse, Figshare, or Zenodo—with the risk of losing visibility in the French landscape, as their interface remains in English.[10]

Finally, it is important to clearly define the scope of a data paper as a data descriptor and not as an analysis paper, or a genealogy of a given project, especially if it is published in a journal that does not have a section dedicated to data papers. In fact, this format is not yet widely used, and many articles whose main goal is to describe a database also include analyses of the data itself, which contributes to the confusion.[11]

The data paper "Mapping the Third Republic" (Gay 2021) provides annual nomenclatures and shapefiles for different administrations of the French Third Republic (1870–1940). As this database can be used both to map localized historical data and to match data from these administrations for statistical analysis, it is aimed at all social sciences with a historical component, especially economic history and historical demography. Hence my choice of the outlet *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, as it is a relatively well-established journal with a wide international circulation and an interdisciplinary readership in the social sciences, with a strong historical component and a predominantly quantitative focus. *Explorations in Economic History* would have been another candidate, but its readership remains limited to economic history. In contrast, the readership of the *Journal of Open Humanities Data* or the *Research Data Journal for the Humanities and Social Sciences* is relatively less focused on cartography or econometric analysis.

In terms of format, the TRF-GIS database is available in the predominant format in the field of economic history: the Stata data format (DTA). However, to make the data accessible to other disciplines, I have also made it available in TXT format, which can be easily imported into any statistical or cartographic software. Finally, I turned to the Harvard Dataverse repository. This repository offers free storage for individual researchers, an ergonomic interface, automatic

[9] A list of journals that publish data papers is provided by Laurence Dedieu (Cirad) at https://doi.org/10.18167/coopist/0057.

[10] There is, to my knowledge, no repository specifically dedicated to the humanities and social sciences, except for Didómena (https://didomena.ehess.fr/), the repository of EHESS' data, which use is limited to its members.

[11] For instance, at the end of an article describing a new database on emancipated slaves in Cape Town in 1834, Ekama *et al.* (2021, 9) analyze the determinants of slave prices using linear regression.

metadata management, and straightforward access for downloading. It is also a repository with a broad international reach. In order to further canvass French-speaking communities, I made sure that the data was distributed on the data.gouv.fr platform as well as in the PROGEDO-ADISP catalog.[12]

Finally, the manuscript of the data paper went through the peer review without much trouble. However, the following comment from an anonymous reviewer clearly shows that this type of article is still misunderstood: "The question is whether this paper constitutes an article. I do not think so. Therefore, I do not think that the review could accept this text, which does not ask a clear historical problem, but is only a kind of guideline and description of a final product."

A Data Paper in the Social Sciences: How?

How to write a data paper so that it is not just a codebook but instead provide a comprehensive understanding of the data produced and proper guidance to their reuse? A current trend is the automatic generation of data papers based on metadata (Schöpfel *et al.* 2019). Although this type of format is much less time-consuming than a standard data paper and allows for rapid dissemination via the automatic harvesting of data catalogs, it does not seem to be adequate for reuse due to the lack of data description in this format. Hence, I suggest avoiding this methodology. Indeed, the description of construction methods and typology choices must be at the heart of the data paper, which only a human can properly achieve.

In this context, the first challenge when writing a data paper is how to structure the text, since it is not quite the same as a traditional article. An excellent template to emulate is the one of the data journal *Scientific Data* of the Nature group.[13] Although this model is aimed at the hard sciences, its structure can be adapted to the humanities and social sciences. The proposed structure begins with a first section ("Context and Summary") that briefly describes the data produced, its scientific context, and its potential reuses. A second section ("Methods") describes in detail all the procedures used in the data production process to ensure reproducibility. Next, a "Data Files" section describes each dataset associated with the data paper. This includes variables, file names, location, formats, and size. A fourth section ("Technical validation") presents the analyses or procedures used to support the validity of the data described—in the humanities and social sciences, this may involve comparison with other sources or with auxiliary data. A fifth section ("Usage notes") allows the author to describe in more detail the procedures for reusing the data and to develop some examples. A final section ("Code availability") describes procedures for accessing the data reproduction code.

A second difficulty for the author of a data paper is the data dissemination. Unlike a traditional research article, where data and reproduction files may be required only at the publication stage, in the case of data papers, they are required when the manuscript is submitted to the journal, with access procedures that must be made explicit in the text. Since the integrity of the data is also assessed by external reviewers, it is necessary for them to have access to the data as early as the review stage. The data must therefore already have a permanent identifier at this stage, such as a DOI, making it possible to attach the data to the data paper and create a "database-data

[12] The repository on data.gouv.fr is available at https: //www.data.gouv.fr/fr/datasets/systeme-dinformation-geographique-de-la-france-de-la-troisieme-republique-1870-1940/. The data is also available in the PROGEDO-ADISP catalogue in the "Historical data" section at http://www.progedo-adisp.fr/enquetes_donhist.php.

[13] Submission guidelines for *Scientific Data* are available at https://www.nature.com/sdata/publish/submission-guidelines. Other templates exist, such as the one offered by *Data in Brief*, available at https://www.elsevier.com/journals/data-in-brief/2352-3409/guide-for-authors. More generally, Kim (2020) provides an overview of the different formats of data papers proposed by various data journals.

paper" ecosystem. However, this can be a problem when reviewers suggest changes for publication. Fortunately, most data repositories allow producers to assign a temporary identifier that is only accessible (anonymously) through a secret code that must be provided to the reviewers.

The structure of the data paper "Mapping the Third Republic" (Gay 2021) follows the *Scientific Data* template described above. After an introduction that presents the scientific context of the database, the body of the paper consists of a "Method" section that explains in detail not only the technical methodology of construction and its limitations, but also the institutional elements underlying the temporal variations of each administration during the Third Republic. For instance, I discuss territorial changes induced by the loss and return of Alsace-Lorraine, the arrondissement reform of 1926, the military reforms of 1873–1874, and the various electoral redistricting laws—these elements are supported by various tables in a 100-page online appendix. This is followed by a description of the fifteen datasets in the TRF-GIS database (variables, storage space, formats, licenses), and a technical validation in two forms: a comparison of the secondary sources mobilized with a set of 175 primary sources (individually listed in the appendix and available as PDFs in the data repository) and a validation of the shapefile construction method by comparison with similar data (the LARHRA 1884 cantons shapefile). The article concludes with a description of where the code and data are located and an example of reuse—specifically, mapping abstention rates in the 1914 general election at the constituency level.

Conclusion

Writing a data paper requires learning various skills as it is a new editorial format in the humanities and social sciences. However, this type of article is growing rapidly and researchers wishing to write a data paper now have access to plenty of feedback through seminars and dedicated conferences. For instance, I could present the data paper describing the TRF-GIS database as part of the "DATA SHS" week organized by the Toulouse University Data Platform (PUD-T) in December 2020, at a webinar dedicated to data papers organized by the CNRS working group "Ateliers de la Données" ("data workshops") in February 2021, at the seminar "From the sources to the GIS" in May 2021, or at the seminar "Histoire et numérique" organized by the History Center Science Po (CHSP) in May 2021, among others.[14]

Many resources are also available from various institutions, such as DoRANum and CoopIST (Cirad), which provide much information on the subject, URFIST or INRAE, which offer training, or the University Data Platforms (PUD) of PROGEDO, which offer seminars on the subject at MSHSs during their annual DATA SHS week.[15]

References

Ali, Nawel Aït and Jean-Pierre Rouch. 2013. "Le 'je suis débordé' de l'enseignant-chercheur." *Temporalités de la recherche* 18: 1–25. https://doi.org/10.4000/temporalites.2632.

Chang, Andrew C. and Phillip Li. 2022. "Is Economics Research Replicable? Sixty Published Papers From Thirteen Journals Say 'Often Not'." *Critical Finance Review* 11: 1–22. http://dx.doi.org/10.1561/104.00000053.

[14] The video of my presentation at the seminar "From the sources to the GIS" is available at https://youtu.be/mBAlRdWR41k, from minutes 2 to 45.

[15] DoRANum's ressources on data papers are available at https://doranum.fr/data-paper-data-journal/. Those from Cirad's CoopIST are available at https://doi.org/10.18167/coopist/0057. A sample of class slides is proposed by the INRAE and available at https://dx.doi.org/10.15454/1.478247389988942E12. See also Reymonet (2017).

Christensen, Garret and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* 56(3): 920–980. https://doi.org/10.1257/jel.20171350.

Cioni, Martina, Giovanni Federico et Michelangelo Vasta. 2020. "The Long-Term Evolution of Economic History: Evidence from the Top Five Field Journals (1927–2017)." *Cliometrica* 14: 1–39. https://doi.org/10.1007/s11698-019-00186-x.

CoSO. 2019. "Pour un politique des données de la recherche : guide stratégique." https://www.ouvrirlascience.fr/pour-une-politique-des-donnees-de-la-recherche-guide-strategique-a-lusage-des-etablissements/.

Desrosières, Alain. 2010 [1993]. *La politique des grands nombres. Histoire de la raison statistique.* Paris: La Découverte. https://doi.org/10.3917/dec.desro.2010.01.

Ekama, Kate, Johan Fourie, Hans Heese and Lisa-Cheree Martin. 2021. "When Cape Slavery Ended: Introducing a New Slave Emancipation Dataset." *Explorations in Economic History* 81: 101390. https://doi.org/10.1016/j.eeh.2021.101390.

Gay, Victor. 2021. "Mapping the Third Republic: A Geographic Information System of France (1870–1940)." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 54(4): 189–207. https://doi.org/10.1080/01615440.2021.1937421.

Gozlan, Clémentine. 2016. "Les sciences humaines et sociales face aux standards d'évaluation de la qualité académique." *Sociologie* 7 (3): 261–280. https://doi.org/10.3917/socio.073.0261.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *Plos Medicine* 2(8): e124. https://doi.org/10.1371/journal.pmed.0020124.

Jiao, Chenyue and Peter T. Darch. 2020. "The Role of the Data Paper in Scholarly Communication." *Proceedings of the Association for Information Science and Technology* 57: e316. https://doi.org/10.1002/pra2.316.

Karial-Cohen, Karine, Claire Lemercier, Isabelle Rosé and Claire Zalc. 2018. "Nouvelles cuisines de l'histoire quantitative." *Annales. Histoire, sciences sociales* 73(4): 771–783. https://doi.org/10.1017/ahss.2019.90.

Khan, Nushrat, Mike Thelwall a,d Kayvan Kousha. 2021. "Measuring the Impact of Biodiversity Datasets: Data Reuse, Citations and Altmetrics." *Scientometrics*, 126: 3621–3639. https://doi.org/10.1007/s11192-021-03890-6.

Kim, Jihyun. 2020. "An Analysis of Data Papers Templates and Guidelines: Types of Contextual Information Described by Data Journals." *Science Editing* 7(1): 16–23. https://doi.org/10.6087/kcse.185.

LARHRA. 2011. *Les cantons français de 1884 à 1966*. Lyon: Laboratoire de Recherche Historique Rhône-Alpes. http://geo-larhra.ish-lyon.cnrs.fr/?q=atlas-historique/territoires-d-etat/evolution-des-cantons-en-france.

Maniadis, Zacharias and Fabio Tufano. 2017. "The Research Reproducibility Crisis and Economics of Science." *The Economic Journal* 127(605): F200–208. https://doi.org/10.1111/ecoj.12526.

Maniadis, Zacharias, Fabio Tufano and John A. List. 2017. "To Replicate or Not To Replicate? Exploring Reproducibility in Economics through the Lens of a Model and a Pilot Study." *The Economic Journal* 127(605): F209–235. https://doi.org/10.1111/ecoj.12527.

Reymonet, Nathalie. 2017. "Améliorer l'exposition des données de la recherche : la publication de data papers." https://archivesic.ccsd.cnrs.fr/sic_01427978.

Robinson-García, Bicolas, Evaristo Jiménez-Contreras and Daniel Torres-Salinas. 2015. "Analyzing Data Citation Practices Using the Data Citation Index." *Journal of the Association for Information Science and Technology* 67(12): 2964–2975. https://doi.org/10.1002/asi.23529.

Ruggles, Steven. 2021. "The Revival of Quantification: Reflections on Old New Histories." *Social Science History* 45 (1): 1–25. https://doi.org/10.1017/ssh.2020.44.

Schöpfel, Joachim, Dominic Farace, Hélène Prost and Antonella Zane. 2019. "Data Papers as a New Form of Knowledge Organization in the Field of Research Data." *Knowledge Organization* 46(8): 622–638. https://doi.org/10.5771/0943-7444-2019-8-622.

Walters, William H. 2020. "Data Journals: Incentivizing Data Access and Documentation Within the Scholarly Communication System." *Insight* 33 (1): 1–18. http://doi.org/10.1629/uksg.510.

Wilkinson, Mark D., *et al.* 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3: 160018. https://doi.org/10.1038/sdata.2016.18.

Zalc, Claire and Claire Lemercier. 2008. *Méthodes quantitatives pour l'historien*. Paris: La Découverte. https://doi.org/10.3917/dec.lemer.2008.01.

Zalc, Claire and Claire Lemercier. 2019. *Quantitative Methods in the Humanities. An Introduction*. Charlottesville (Virg.): University of Virginia Press. https://doi.org/10.2307/j.ctvbqs963.