Accelerating social science knowledge production with the coordinated open-source model

Turek, Konrad

DOI:
10.31219/osf.io/3brjv

Publication date:
2023

Document Version
Early version, also known as pre-print

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
Turek, K. (2023). *Accelerating social science knowledge production with the coordinated open-source model*. OSF. https://doi.org/10.31219/osf.io/3brjv

Accelerating Social Science Knowledge Production with the Coordinated Open-Source Model

Konrad Turek

Tilburg University, Human Resource Studies Department &
Interdisciplinary Herbert Simon Research Institute
Contact: k.l.turek@tilburguniversity.edu

Preprint v.1.2 (early draft), published 06/04/2023

Please note this is an ongoing work. Fell free to contact the author with the comments!

Abstract

With the growing complexity of knowledge production, social science must accelerate and open up to maintain explanatory power and responsiveness. This requires redesigning the front-end of research to build an open and expandable knowledge infrastructure that stimulates broad collaborations, enables breaking down inertia and path dependencies of conventional approaches, and boosts discovery and innovation. This article discusses the coordinated open-source model as a promising organisational scheme that can supplement conventional research infrastructure in certain areas. The model offers flexibility, decentralization, and community-based development and aligns with open science ideas, such as reproducibility and transparency. Similar solutions have been successfully applied in natural science, but social science lags behind. I present the model's design, and consider its potential and limitations (e.g., regarding development, sustainability and coordination). I also discuss open-source applications in various areas, including the first open-source survey harmonization project in social science.

Keywords

Open Science, Open Source, Reproducibility, Innovation, Path Dependency

Introduction

In times of rapid social change and increasing complexity of knowledge production, social science must adapt its collaborative and management processes to be able to accelerate and advance (Gerring, Mahoney & Elman 2020; Hofman et al. 2021; King 1995; Vazire 2018). The explanatory power and responsiveness of science are increasingly dependent on the ability to utilise diverse resources, build broad collaborations, and coordinate shared work (Sterman & Wittenberg 1999). Recent decades have shown that social science is often surprisingly slow in responding to societal challenges, such as pandemics and rising inequalities (Besancon et al. 2021; Hirschman 2021; Savage 2021). It is also lagging behind natural science in adopting novel organisational schemes, such as crowd-based collaboration or open-source model, that offer an interesting alternative to more conventional approaches (Franzoni & Sauermann 2014; Moshontz et al. 2018). Novel challenges appear, for example, in artificial intelligence (e.g., language models, such as ChatGPT) which creates still unexplored potentials and risks for the knowledge production processes. Social science must develop novel solutions to overcome the limitations of conventional knowledge infrastructures: the self-limiting and path-dependent mechanisms that tend to reproduce certain methods of knowledge production and limit flexibility and innovation (Benbya, Jacucci & McKelvey 2006; Hirschman 2021).

The current debate on open science (Altman & Cohen 2022; Fecher & Wagner 2016; Freese, Rauf & Voelkel 2022; Friesike et al. 2014) provides a good moment to discuss these challenges and consider new approaches to scientific cooperation and knowledge production. The response to the pandemic showed that science could be dynamic and integrate global-scale efforts (Altman & Cohen 2022; Callaway 2020). However, a system of tools, institutions and incentives is needed to make fast and open collaboration a standard approach to producing knowledge. The existing links and dependencies provide a natural fundament for creating a common infrastructure to share resources and coordinate efforts. Most of all, a model for such open and broad collaboration exists and functions well in many areas – this is an open-source model.

In this article, I present the coordinated open-source model as a promising and intriguing organisational scheme that can help opening the knowledge infrastructure in social science. First, it offers flexibility, decentralised control, and community-based development, which enables breaking down inertia and path dependencies, and opens possibilities for novel research questions and applications. Second, it can allow social science to better identify and respond to societal challenges, e.g., by improving access to and the processing speed of high-quality data. Third, the open-source model aligns with the priorities of the open science movement and emanates the fundamental idea of science as a collaborative process, where researchers benefit from sharing their efforts and contribute to faster and more ambitious scientific progress. As a result, a crowd-based organisational scheme can be supplementary to the more conventional research infrastructure in certain areas, helping advance and accelerate social science.

Although open-source and crowd-based collaboration is attracting growing attention from the academic community, it rarely appears in social science. To introduce it and place it in the social science landscape, I will first discuss how opening the knowledge ecosystem can stimulate scientific development. Then, I will present the design, potential, and limitations of the open-source model in social science knowledge production. As a case study, I will refer to the first open-source survey harmonisation project, the Comparative Panel File (Turek, Kalmijn & Leopold 2021), and evaluate the first two years of this experimental initiative. Finally, I will consider other potential areas of implementation.

Opening the knowledge ecosystem

The trend towards open and transparent science is changing how knowledge is produced, verified and distributed (Altman & Cohen 2022; Callaway 2020). At the fundamental level, "opening" science means making knowledge more transparent, accessible, reproducible, and reliable. Many relatively simple solutions and practices that support these goals have been popularised, e.g., open-access publishing, preprints, preregistration, replications, and transparency in data management (Altman & Cohen 2022; Firebaugh 2007; Freese et al. 2022; Nosek et al. 2015). During the last decade, we also witnessed large-scale initiatives aimed at reforming the more broadly defined scholarly knowledge infrastructure (Altman & Cohen 2022; Edwards et al. 2013), such as building institutional networks, open data infrastructures and data hubs, such as Open Data Infrastructure for Social Science and Economic Innovations (Odissei) in the Netherlands, Dataverse Network, and openICPSR (Freese & King 2018; Gerring et al. 2020; Kapiszewski & Karcher 2020; King 2007; Moshontz et al. 2018). The importance of these trends toward open science has also been recognised by the European Commission that in 2016 released the general recommendation for establishing a European Open Science Cloud aimed to support open science and open innovation, particularly in the area of sharing and reuse of scientific data (EU 2016). However, despite many advancements, social science knowledge infrastructure is still limited by deeper structural problems that challenge two major goals of science: the ability to explain social reality and respond to novel problems.

Explanatory power

The explanatory power of science refers to the ability to describe the observed phenomenon using theory, reveal mechanisms or relations underlying observed data, and predict generalised conclusions for new situations (Franck 2002; Sterman & Wittenberg 1999). The explanatory power of social science is challenged by the increasing specialisation and complexity of scientific problems. To understand contemporary societies, researchers increasingly apply advanced and sophisticated theories that include causal, multilevel and longitudinal mechanisms (Bak-Coleman et al. 2021; Brüderl, Kratz & Bauer 2019; Crossley 2021; Gangl 2010; Gerring et al. 2020). This process is accompanied by methodological, statistical and computational advancements. They steadily increase the demands for diverse analytical and programming skills, as well as high-quality data allowing verification of the increasingly complex theoretical hypotheses (Edelmann et al. 2020; Hofman, Sharma & Watts 2017; Jones 2009; Salganik 2017).

For example, more complex theoretical and empirical approaches are required in the increasingly popular life-course studies, which focus on continuity and change across individuals' lives and link it to contextual factors (Piccarreta & Studer 2019). Scholars also complained at growing complexity of organisational research, a multidisciplinary field focusing on multilevel and dynamic relations between individuals and their organisational environment. Specifically, the process was related to increasing number of unreliable research, growing gap between theory and practice, and a lack of cumulative progress (Anderson 1999; Starbuck 2006). Similar concerns have been expressed in ecological research (Low-Décarie, Chivers & Granados 2014), psychology (Vazire 2018), or epidemiology (Galea, Riddle & Kaplan 2010), and general science (Anderson 1972; Evans & Foster 2011; Franck 2002; Wuchty, Jones & Uzzi 2007).

Responsiveness

The second major challenge of social science concerns its slow responsiveness (Savage 2021; Starbuck 2006; Van De Ven & Johnson 2006). Responsiveness refers to the ability to identify and address societal challenges and produces outcomes practically relevant in dealing with the challenge. As such, responsiveness constitutes the basis for the speed of reaction and knowledge production, applicability of this knowledge, and openness to new perspectives and approaches.

The sudden and profound strike of the Covid-19 pandemic elucidated how important it is to quickly produce reliable and innovative scientific solutions (Dahlander, Gann & Wallin 2021; Kühne et al. 2020). The conventional, institutionalised academia is limited in this regard, for example by its slow research and review process. Nevertheless, the scientific community responded to the pandemic unprecedentedly with solutions that made the knowledge ecosystem more open, networked and responsive (Besancon et al. 2021; Callaway 2020; Lucas-Dominguez et al. 2021). Extensive data sharing, resources and efforts helped address the common challenge more efficiently. Preprints with replication files allowed quick crowd-based review (Fraser et al. 2021; Watson 2022). Even in case of problems and errors, many correction tools worked well, e.g., unreliable studies have been identified and retracted. Briefly, the opening of the knowledge production speeded up investigations, publishing process, verification of findings and development of solutions.

Limitations of social science: the case of inequality studies

Several scholars have recently discussed the problems of weak explanatory power and slow responsiveness of social science, taking the history of inequality studies as an example (Hirschman 2021; Jackson 2022; Savage 2021). This debate was stimulated by the late recognition of the sharply rising inequalities currently placed among the major challenges for developed societies. Although income and wealth disparities have been steadily growing since the 1980s, the trend was recognised only in the early 2000s. In their impactful works, Piketty and Saez (2003) and Piketty (2014) showed a progressing accumulation of incomes among the richest. These results surprised the world and reframed the public debate to focus on the 'rich as a social problem' (Savage, 2021, p. 3), even provoking social movements such as Occupy Wall Street and large-scale riots.

Given the importance of this problem, Hirschman (2021) wonders why it took economists as long as two decades to identify and incorporate this problem into the public debate. Based on an in-depth analysis of the history of economic studies on inequality, he blames the fundamental limitation of the conventional system of knowledge production. As he suggests, once established, the ways, tools, norms and perspectives of a specific knowledge infrastructure make certain research questions, applications and outcomes more "doable". Consequently, researchers follow these established paths, and those who succeed and gain more power benefit from maintaining the system. However, this path dependency can constrain novelty and narrow the research perspectives, and consequently, it limits the responsiveness and explanatory power of social science.

Piketty and Saez's success in attracting public attention has to be ascribed mainly to their novel take on the problem of rising inequalities (Savage, 2021). Instead of an abstract Gini coefficient, inequalities have been visualised by targeting the specific top-income group (e.g., top 10% or 1%), making it concrete and easy to understand. The analyses also covered a long historical perspective, reaching back to the early twentieth century. However, as Hirschman (2021) notes, these elements did not guarantee success. Although the authors used openly available data and relatively simple indicators, their approach and broad scope were unusual for the economics of that day yet not entirely novel. Similar methods were used in the first half of the twentieth century but were largely abandoned since they did not fit the dominant economic knowledge infrastructure of the late 20th century. Also, increases in top incomes have occasionally been spotted, yet they have not attracted systematic public attention. According to Hirschman (2021), what was necessary for success was the development of a new knowledge infrastructure. Piketty and colleagues redesigned the inequality data infrastructure using existing data sources and shared it with the scientific community. It allowed monitoring of the trends, stimulated further systematic research and influenced the public debate.

There are more lessons to be learned from this story. For decades, the single economic paradigm and income-related outcomes have strongly dominated research on inequalities. Jackson (2022) argues that this domination had adverse policy effects in the United States because it narrowed the size and scope of policy interest and hampered open and innovative reforms. It has also marginalised alternative approaches in academic and public debates (DiPrete & Fox-Williams 2021; Jackson 2022). For example, much less attention has been paid to other aspects of inequalities in contemporary societies, such as life-course accumulation mechanisms (Crystal & Shea 1990), opportunity-based approaches (Sen 1993), and relational approaches (Tomaskovic-Devey & Avent-Holt 2019). Recently, the pandemic reminded us about the multidimensionality of inequalities, which also include disparities in health, access to care and education, work quality, and secure living environment, which cannot be explained by economic stratification only (Blundell et al. 2022).

An example of such a neglected perspective is the organisation-oriented research on inequalities. In the late 1970s, Baron & Bielby (1980) considered organisations a crucial area of inequality development, yet this idea received a broader interest only four decades later (Amis, Mair & Munir 2020; Riaz 2015; Tomaskovic-Devey & Avent-Holt 2019). For example, recent studies took an organisational lens to examine racial (Ray 2019) and gender (Acker 2016) inequalities. Starbuck (2006) argues that one of the primary reasons for this slowdown in organisational research on inequality was its multidisciplinary character, which was much more problematic to fit into conventional academic frameworks than the more commonly acceptable macroeconomic paradigm. He claims that institutionalised way of producing knowledge in organisational science hampers theoretical progress and constrains practical value. Another vital aspect that affected the development was a lack of sufficient organisational data, e.g., organisation-employee linked data, which are usually much more demanding to collect (Tomaskovic-Devey & Avent-Holt 2019). Jackson (2022) also argues that the strong position of economists in the United States is partly due to their better access to high-quality administrative data based on well-developed networks and connections to government gatekeepers.

What blocks explanatory power and responsiveness?

The history of inequality studies reveals the importance of reconsidering how social science operates. Three points appear critical. First, without appropriate data infrastructure, social scientists are left with unverified theories and little legitimisation for suggesting practical solutions. Sharing and integrating data has improved the explanatory power and responsiveness of many areas of natural science, yet social science is lagging behind (Firebaugh 2007; Freese 2007; Gerring et al. 2020; Vazire 2018). Better access to high-quality data is essential to advance knowledge (e.g., the case of organisation-approach to inequalities). Hirschman argues that data infrastructure can drive the ignorance loops: "gaps in the infrastructure shape what research is done, the research that is done shapes priorities for maintaining and creating the infrastructure" (Hirschman, 2021: 780). For example, he points out that systematically small numbers of minority respondents in US surveys created a gap in knowledge infrastructure that significantly constrained research on minorities. Moreover, access to data and data-sharing practices can be crucial for responding to rapid societal challenges. In the case of the COVID-19 pandemic, open data sharing was a fundament for efficient scientific cooperation, quick verification of the results and implementation of life-saving solutions (Besancon et al. 2021; Watson 2022). Studying such collectively experienced crises can be facilitated by comparative longitudinal survey data – although they exist worldwide, their harmonisation is laborious and demanding. Various harmonisation initiatives aimed to speed-up acquisition and management of such data (Dubrow & Tomescu-Dubrow 2016), however, they are still not fast and flexible enough to meet researchers' expectations (Turek et al., 2021). To keep up with the growing complexity of knowledge production, "social scientists need to continue to build a common, open-source, collaborative infrastructure that makes data analysis and sharing easy" (King, 2011, p. 720). In short, an open and expandable data infrastructure, with broad and fast access to high-quality data, is one of the fundaments for the explanatory power and responsiveness of science.

Second, extensive, international and multidisciplinary collaborations are increasingly important for successful scientific research (Woolley et al. 2015; Wuchty et al. 2007; Zuo & Zhao 2018). Contemporary science is strongly networked and interconnected at many levels, with overlapping goals and data resources. However, social researchers prefer smaller research teams and rarely attempt to coordinate activities or share the work (Auspurg & Brüderl 2021; Hucka et al. 2015). Although broader collaborations are not always required, they might improve the speed and efficiency of knowledge production in many areas. Field-specific research lenses can constrain more general knowledge advancement – facilitating certain research topics and approaches while simultaneously suppressing others. Narrow theoretical perspectives also limit contribution to public debate and practical impact of research because they may ignore some important aspects of societal problems, as in the case of the dominance of the economic perspective in inequality studies (Jackson, 2022). We need efficient infrastructure for managing collaborations and sharing work. *Ad hoc* and occasional efforts do not guarantee a fast and cumulative scientific process. In particular, the increasingly computational nature of social science requires new platforms for data management and new ways of interdisciplinary team collaboration (Lazer et al. 2020).

Third, conventional knowledge infrastructure tends to be conservative and self-limiting, constraining flexibility and innovation required to identify and address novel research problems. Hirschman (2021) states that problems recognised in the economic approach to inequalities are inherent to each conventional knowledge infrastructure. In more general terms, these concerns refer to path dependency mechanisms that have been recognised to limit research and development activities (Benbya et al. 2006; Coombs & Hull 1998; Volberda, Schneidmuller & Zadeh 2021). Path dependency is a historical development trajectory in which past decisions shape and constrain present decisions, even if contextual factors have changed and alternative decisions could be better (David 2007). It develops well-established patterns of behaviours, routines, know-how resources and petrifies dependencies between system elements. Such solutions have functional benefits because they are often efficient, reduce costs or effort, increase returns and stabilise relationships in the system. However, path-dependent systems are limited in novel situations and innovations because they reinforce behaviours consistent with prior developments (Arthur 1994; Cohen & Levinthal 1990). Academic systems seem to be perfect examples of such institutional path dependencies (Hollingsworth 2008; Krücken 2003). Academia is still largely structured by ideas and infrastructures forged in the nineteen and early twenty century, keeping scientific activity firmly within the boundaries of universities, research institutes and companies (Franzoni & Sauermann 2014; Savage 2021). The inertia of the scholarly system produce systematic ignorance that limits novelty. “Past priorities shape existing knowledge infrastructures that in turn channel researcher attention toward some problems and away from others” (Hirschman, 2021, p. 742). Knowledge infrastructure should provide enough room for discovery and innovation and enable breaking path dependencies (Swedberg 2020).

As the pressure to maintain explanatory power and practical applicability increases, social science must accelerate and open knowledge production. Meeting this challenge goes, however, beyond the back-end practices regarding dissemination and verification of results, which are increasingly popularised by the open science movement. As Friesike et al. (2015) argue, we must redesign the front-end of the research and innovation processes. Similarly, Arthur and Cohen (2022, p. 2) call to "entirely re-engineer the systems of scholarly knowledge creation, dissemination, and discovery". We must build a stable knowledge infrastructure that increases access to high-quality data, stimulates broad collaborations, and is open to discovery and innovation. Such infrastructure must provide functional components that are modular, interoperable and reusable (Almaatouq et al. 2021). Despite being considered the cradle of liberal and progressive thinking, social science is 'remarkably conservative' in its academic practice (Savage, 2021, p. 7), especially with regard to strong disciplinary boundaries and limited cooperation and communication. Narrow and closed knowledge infrastructure limits the ways and priorities for scientific knowledge production and constraints science's explanatory power and responsiveness.

The rise and progress of open-source initiatives

In this light, the open-source model (also called crowd-based or networked collaboration) should be considered a promising organisational scheme for producing knowledge. The underlying idea is to share the work in open networks of contributors and share the outcomes with a broader community of users. Active participation by contributing to the initiative is voluntary and unpaid. Usage and application of the outcomes are free of charge and do not require any active contributions. Open-source results are public goods whose economic value is derived from collective potential, challenging the capital-based ownership model.

Initially, open-source cooperation was implemented in software development. The idea was proposed already in the 1980s as an alternative method of software development, but it gained more attention in the early 2000s with advancements in computer technologies and programming frameworks. Today, Free/Libre Open-Source Software (FLOSS) is a widespread programming solution (Crowston et al. 2012). However, the open-source idea goes beyond software development and can be found in various virtual collaborations that aim to generate ideas, knowledge, and solutions by involving a large number of external actors. Such collaborations are organised around virtual platforms that connect persons, processes, interfaces, services, and artefacts. By integrating dispersed knowledge, resources and opportunities, virtual collaborations can maximise capabilities existing in the broader ecosystem and co-create value and innovations (Abbate et al. 2021; Esposito De Falco et al. 2017).

Open-source projects usually involve several groups. As in the traditional business model, the core-development team can initiate development and control critical activities. However, the team's composition, role and authority are much more fluid – they arise from bottom-up processes, e.g., as a result of the contributions to the commonly agreed goal, and can vastly differ between projects and change over time (Bonaccorsi & Rossi 2003). Moreover, a critical role in open-source projects is played by peripheral developers, who temporarily and voluntarily contribute to the product (Crowston et al. 2012; Setia et al. 2012). Although peripheral developers usually have shorter affiliation with the project than core developers, in many open-source initiatives (especially in the more mature stages of product development), they contribute much to the success. Both of these groups have to be separated from the group of passive users (who apply the product yet do not actively contribute to its development) and active users (they may report errors, comment or request features).

Adaptation of the open-source model in science has been relatively slow and selective. Open-source scientific initiatives began to appear at a larger scale in the early 2000s, and since then, the amount of research done this way has steadily risen, but mainly in natural sciences, e.g., biology, medicine, ecology, physics and geography (Franzoni, Poetz & Sauermann 2021; Hucka et al. 2015; Kullenberg & Kasperowski 2016; Pfaff & Hasan 2007). For example, open-source solutions contributed to biomedicine (Rai 2005), computational chemistry (Lehtola & Karttunen 2022), and general statistical tools (e.g., Python and Stan programming languages or R statistical package). Open-source cooperation still rarely appears in social science (Beck et al. 2022; Franzoni & Sauermann 2014; Friesike et al. 2014). When searching for "open source" in the Web of Science portal archives, we find a continuously rising trend in the number of publications. However, most of the findings come from computer sciences and engineering journals (due to the technical nature of this issue), and the rest is dominated by natural science, such as biology, ecology, astronomy, or physics. When it comes to broadly defined social science and humanities, Web of Science query sums it up to merely 1% of all entries (less than 15.000 from the total of 150.000).

The promise of an open-source model in social science

The open-source model is a promising alternative to conventional ways of cooperation and production of social knowledge for several reasons. First, although as an organisational scheme, it is relatively new, it builds on ideas with a long history in the scientific debate. Specifically, the model answers the call for equal, inclusive and open communication expressed by the 20th-century philosophers (Breznau 2021). For example, Jurgen Habermas (1984) considered open communication as a solution for inequalities in production and consumption of communication. The open-source model also embodies the vision of science as a collaborative and cumulative process. Much of modern research not only exceeds the capabilities of a single researcher but also requires broader and more open cooperation. Researchers benefit from sharing their efforts and contribute to faster and more ambitious scientific progress. As expressed by Karl Popper (1959 [1934]), this is a never-ending, always incomplete process focused on temporary solutions and elimination of errors. Open cooperation also opens ways to innovation. For example, Charles S. Peirce (1902) emphasised that the context of discovery stimulates open-minded approaches and unbiased conceptual frameworks necessary for explaining reality. Eventually, Robert K. Merton (1973 [1942]) famously argued for the need for communalism, universalism and organised scepticism, which is often considered a fundament for the contemporary open science movement.

These ideas fit well today's knowledge creation which has become a collaborative enterprise, strongly dependent on virtual research cooperation and distributed research networks (Almaatouq et al. 2021; Aydinoglu 2013; Wuchty et al. 2007). Open-source collaboration allows for a large and diverse base of contributors, potentially expanding the range of scientific problems that can be addressed. The heterogeneity of actors also amplifies collective intelligence and creativity (Arza et al. 2018). Crowd-based projects are not only a way to accelerate the research process by sharing tasks but are very often an essential requirement to conduct a large-scale project that exceeds the possibilities of any single team. Such collaborations are much better adapted to deal with many modern scientific problems, particularly those more complex, interdisciplinary, and heavily dependent on computer technologies and dispersed knowledge (Felin & Zenger 2014). Virtual teams with well-developed communication systems and distributed division of labour create so-called transactive memory systems (Chen et al. 2013). In such systems, information and knowledge are allocated, stored, and retrieved collectively. Cohen and Levinthal (1990) define this as an absorptive capacity – an ability to identify, assimilate, and exploit knowledge from the environment. By doing this, virtual teams can deal with complex tasks and achieve higher performance.

A good illustration of the communal approach to science is crowd research, also called crowd science, networked science, or crowdsourcing research (Auspurg & Brüderl 2021; Beck et al. 2022; Franzoni & Sauermann 2014; Uhlmann et al. 2019). Here, researchers who are not formally linked cooperate in an open network to investigate the same research question or work on the same data. For example, Salganik et al. (2020) used scientific mass collaboration (160 teams) to perform the same research task – measuring the predictability of specific life outcomes using the same data but various methods. SCORE (Alipourfard et al. 2021) is a crowdsourced empirical project to assess the credibility of results published in social and behavioural science by engaging hundredths of researchers in distributed tasks, such as reproduction and replications. Psychological Science Accelerator is an example of a distributed network of laboratories designed to enable and support crowdsourced research projects (Moshontz et al. 2018).

However, to separate from crowd research, scientific open-source initiatives should be narrowed down to those focused on a bottom-up co-development of research tools and infrastructures. Open and flexible infrastructures can enable faster responses to unpredictable challenges with novel ideas and solutions (Aydinoglu 2013). By exploiting extended knowledge and resources, virtual collaborations can increase performance (Volberda et al. 2021). They may also pursue open innovation by allowing unconstrained inflow and outflow of knowledge to accelerate value creation and build new applications (Chesbrough 2003). Studies show that an environment with open governance and interdisciplinary and diverse teams is more likely to generate innovative outcomes (Dahlander & Gann 2010; Felin & Zenger 2014) and high-impact scientific publications (Banal-Estañol, Macho-Stadler & Pérez-Castrillo 2019).

Furthermore, open-source initiatives can accelerate research, increase efficiency and stimulate the accumulation of knowledge. For example, the code designed for openly available data can be reused by other scientists, decreasing the time and costs of research. Importantly, it can also limit barriers to initiating and conducting studies, especially those riskier and with high entry costs (Arza et al. 2018; Franzoni & Sauermann 2014; Jones 2009). Re-usage of the code also allows for spotting problems or errors and faster correction. Overall, it aligns with the open-science principles of transparency and reproducibility.

Another important aspect is the relatively low cost of development and management of open-source infrastructure. As a result, open-source projects can yield much higher returns to investment than conventional approaches, making it an attractive option for the mostly-underinvested science. A broad scientific community organised around a crowd initiative can also have a stronger position in seeking financial funding (Hucka et al. 2015).

Finally, the open-source model can enrich the knowledge ecosystem with new capacities. Knowledge ecosystem, in this perspective, refers to a network of actors and institutions (e.g., research organisations, universities, and for-profit innovators) that share knowledge and generate new solutions and technologies through joint research work (Abbate et al. 2021). Primarily, it consists of conventional knowledge infrastructures that focus on knowledge exchange. Recently, various institutional reforms aimed at opening science through institutional networks, open data infrastructures and data hubs largely contribute to opening so-defined knowledge ecosystems. However, Abbate et al. (2022) argue that virtual collaborations, such as open-source initiatives, belong to a qualitatively different form of a 'capability' ecosystem. The capability ecosystem allows going beyond the exchange of knowledge and systematically stimulates the generation of new capabilities. Here, open innovation moves upfront as the major goal, although unpredictable and achieved in uncoordinated ways. In order to transition from the knowledge exchange to the capacity-generating framework, the ecosystem must be open and capable of breaking path dependencies. This can be done through bottom-up, self-reinforcing, and non-deterministic mechanisms. Drawing upon the complexity theory (Benbya et al. 2006; Elder-Vass 2010), we can see the open-source model as a complex adaptive system capable of self-organisation and generating emergent properties, i.e., qualitatively novel outcomes that develop in uncoordinated interactions and are irreducible to the inputs. As Sterman & Wittenberg (1999) argue, openness, dynamics, and adaptability are essential for developing and expanding new paradigms, allowing scientific revolutions, and enhancing the explanatory power of science. For example, the recent advances in Artificial Intelligence (AI) language models (e.g., ChatGPT) can open new and unexpected ways for code-based collaborations. Already now, such tools perform well in writing and translating code, designing algorithms or preparing code-based documentation. It is difficult to asses the consequences of the AI-revolution for research processes, but the open-source model seems to be one the ways for harnessing the potentials of these developments.

Potential fields of application

Although the open-source model is still rare in social science compared to more conventional cooperation methods, it is not absent. This section will point to several promising areas for implementing coordinated open-source projects.

Computational social science

Computational advancements are a major driving force behind the use of open-source and crowd cooperation in social science. Analysis of social phenomena increasingly often applies various computer-based methods, such as machine learning techniques, simulations, natural language processing, data mining, network analysis, and automated text analysis (Edelmann et al. 2020; Hofman et al. 2021; Lazer et al. 2020; Salganik 2017). They are dependent on complex programming code and code-based cooperation. Very often, the general programming framework can serve multiple purposes beyond a single project. Programming components and sets of solutions developed for a particular goal can be reused and adjusted by other teams, speeding up the research processes.

A good example is Agent-Based Modelling (ABM), a method that simulates adaptive behaviours of agents (e.g., individuals) who influence one another and react to the environment (Macy & Willer 2002; Steinbacher et al. 2021). ABM allows studying collective behaviours and emergence of system structures, and experimenting with hypothesised mechanisms. Although ABM is gaining attention in social science, it is rarely applied. One of the major obstacles is the technical complexity – ABM programming requires skills and time. Therefore, several open-source platforms emerged to share and reuse ABM code and modelling frameworks (Devillers et al. 2010; Janssen et al. 2008; Marwick 2016). For example, CoMSES Net, the Network for Computational Modeling in Social and Ecological Sciences (www.comses.net), is an open community of researchers interested in ABM of social and ecological systems.

Crowd-based experiments and virtual laboratories

Another promising field is virtual collaboration via digital platforms (Esposito De Falco et al. 2017). In particular virtual, crowd-based experimentation labs are gaining much interest (Beck et al. 2022; Hofman et al. 2021; Horton, Rand & Zeckhauser 2011; Mason & Watts 2012; Salganik, Dodds & Watts 2006). They allow collecting experimental data at a scale and pace not available in physical laboratories. A similar idea can be recognized in wiki surveys proposed by Salganik & Levy (2015) as an open and crowd-based survey instrument in which respondents can also be contributors. Specifically, respondents' answers to open questions are added to the list for further participants. The authors show that such a collaborative and adaptive design can help generate and evaluate ideas. To fully utilize the opportunities of the digital world, Almaatouq et al. (2021) suggest developing a broader virtual lab infrastructure designed as an open, flexible and modular systems, where the research community can easily adapt the technical solutions to run larger, faster and more complex experiments. They argue that the new open experimental ecosystems can boost creativity, leading to new types of methods and theories unavailable with conventional approaches.

Open-source code for secondary data analysis

The open-source model also offers much to secondary data analysis (SDA). SDA, where researchers use existing data collected by others, is very popular in social science, particularly for large and costly population surveys. Quite often, preparation of the data for analysis (e.g., combining files, cleaning the data, integrating and harmonising separate surveys) proceeds with a similar workflow regardless of the research topic. Thus, it can be shared to increase efficiency (Fecher et al. 2015). Moreover, social science observes a growing interest in register and administrative data, which are available in more countries (Connelly et al. 2016). Preparing and managing such data is technically challenging, so code sharing (if allowed by security protocols) can have many advantages. One example of an open-source SDA initiative is Gateway to Global Aging Data platform (www.g2aging.org). It provides free resources for harmonizing survey data on ageing-related issues and encourages research collaboration and data sharing (Jain, Min & Lee 2016). Furthermore, some secondary survey data sources include users' code repositories that enrich usability and applications (e.g., UK Understanding Society Household Longitudinal Study). In many other cases, researchers share such code directly, e.g., at GitHub code repository or private websites.

Challenges and the coordinated open-source model

Open-source data infrastructure can contribute to advancing and accelerating interdisciplinary social science, but several challenges must be addressed. Studies on open-source projects over the past few decades show that abandonment and termination of such initiatives are not uncommon. I will discuss that two major challenges for such initiatives, namely the stimulation of crowd-based development and assurance of long-term sustainability, and argue that some level of coordination is required in open-source scientific initiatives.

Crowd-based contribution and development

In science, sharing the analytical code for publicly available secondary data sources is commonly perceived as positive for stimulating and advancing research, however, active participation in such practices is less popular (Fecher et al. 2015; Linåker & Regnell 2020; Scheliga et al. 2018). Freese (2007) admits there are good reasons for researchers not to do it, as they may have spent much time writing the code and can be reluctant to allow others to benefit from their work. Gaining experience working with a particular dataset may also be seen as a competitive advantage for a researcher or team if they plan to use the same code for future projects. The pressure of the competition is accompanied by the lack of direct stimuli to contribute with voluntary work for the broader community. Communication always has some cost (Baldwin & Clark 2006). “Too much effort!” was the foremost researcher’s response to why they do not share data and code in a survey conducted by Fecher et al. (2015) among the users of the German Socio-Economic Panel (SOEP) survey. Additionally, researchers can fear errors in their code can be found and negatively affect their careers. Briefly, efforts required to share the code and perceived risks often outweigh the potential individual benefits.

This type of situation is well known to social scientists as the collective action problem or free-riding problem (Baldwin & Clark 2006; Olson 1965). Although it would benefit everyone in a community to cooperate, individualistic interests and conflicting goals often discourage collective efforts. Open-source collaborations have sought two kinds of solutions to this problem: organisational and technical. The first is supporting community building and active involvement of peripheral developers (Bonaccorsi & Rossi 2003; Fecher et al. 2015; Franzoni & Sauermann 2014; Matei & Irimia 2014; Shah 2006). Such actions primarily aim at stimulating intrinsic motivation (e.g., by gamification, supporting community commitment, reputation or reciprocity norm), rising interest (e.g., by providing access to materials or outcomes), or facilitating formal recognition (e.g., citations or increasing career prospects). For instance, Avelino et al. (2019) investigated many GitHub open-source projects and concluded that personal and professional needs were the primary motivations to contribute. In a systematic literature review on the barriers to contributing to open-source software projects, Steinmacher et al. (2015) find that newcomers are often discouraged by a lack of social interactions with project members that would enable better socialisation and identification with the initiative.

Technical solutions to the collective action problem are based on the idea that successful development and active collaboration in open-source projects depend on technical design and the quality of the infrastructure (Freese & King 2018). Reluctance to transparency initiatives and code and data sharing may result from design flaws in the knowledge ecosystem (Gerring et al. 2020). For example, Baldwin and Clark (2006) argue that a mode modular and flexible codebase architecture can stimulate contributors' engagement and mitigate the free-riding problem. Streamlining the bottom-up processes is usually done by microtask workflows that modularise and pre-specify goals and actions (Valentine et al. 2017). Also, Avelino et al. (2017) also emphasise the importance of technological solutions that make contributing easy and support continuous integration.

Sustainability and coordination

The second biggest challenge for open-source projects is their long-term sustainability. Even though design, functionality, and community are important, they do not guarantee that the initiative will continue. Contrary to conventional projects, financing plays here a smaller (though not negligible) role. A key sustainability factor that is less often considered is coordination. The idea of "coordination" of open-source collaborations is not straightforward because decision processes have predominantly bottom-up, self-organising and decentralised character (Bonaccorsi & Rossi 2003; Setia et al. 2012). Nevertheless, also open-source projects require some organisational structure, management framework and leadership. While a rigid management style may harm collaboration and limit the unique values of virtual research collaborations, some general leadership is required (Aydinoglu 2013; Duparc et al. 2022; Felin & Zenger 2014; Matei & Irimia 2014; Volberda et al. 2021). Studies on information and knowledge-based systems suggest that if the growing complexity of the environment is not managed appropriately, such systems fail (Benbya et al. 2006). Ongoing coordination of the core management processes is especially vital for scientific applications, where the contributors base is relatively limited, expertise is dispersed and diverse, and links to the academic environment are complex.

Successful coordination of complex adaptive systems, such as open-source crowd-science initiatives, requires integrating top-down and bottom-up processes. First, it should ensure the daily functioning of the open-source platform, e.g., resolving technical issues. A core management team does not equate with the group of 'core developers', yet it can initiate, stimulate and structure the flow of contributions (Bonaccorsi & Rossi 2003). It may provide positive feedback to bottom-up, self-organising developments, facilitating new functionalities and structures (Duparc et al. 2022). For example, stronger coordination can be required in more complex, open-ended tasks, where pre-specification and modularisation are difficult (Valentine et al. 2017). The success of scientific open-source initiatives largely depends on integrating the model with the broader scientific ecosystem through active promotion, strategic collaboration with key actors, or acquisition of competitive grants. Thus, top-down management can help navigate the environment and make strategic moves.

Another vital issue is expertise coordination. In open-source initiatives, the expertise is dispersed and diverse, which may cause problems and even contribute to the failure of open-source initiatives (Faraj & Sproull 2000; Pfaff & Hasan 2007). For example, Nupedia – the predecessor to Wikipedia – was abandoned primarily due to the complicated review process of articles (Rosenzweig, 2006). Wikipedia's success was based on its novel and open knowledge management system that reduced the review and edition time. Expertise coordination, especially in scientific applications, also relates to the quality and credibility of outcomes (Franzoni & Sauermann 2014; Friesike et al. 2014). Crowd science must still meet rigorous scientific standards. Scheliga et al. (2018) found that quality assurance and feedback mechanisms were pivotal aspects of the success of crowd science projects.

The topic of coordination (governance or management) of open-source initiatives has been gaining increasing interest from management literature as a novel process that is intriguing and challenging to define. As Volberda et al. (2021) note, the theoretical perspective on managing knowledge development and stimulating innovations in contemporary societies is shifting from static models to more dynamic theories. The latter (e.g., managerial agency theories) emphasise the active role of managers in shaping strategists in knowledge development and innovation. In general, management literature (Duparc et al. 2022; Felin & Zenger 2014) agrees that the more traditional, closed governance (based on authority, property rights, and strong hierarchies) is less beneficial for open innovation and knowledge exchange than open governance (build upon a larger number of external linkages). However, the idea of open governance is broad and unclear. For example, Felin & Zenger (2014) consider three forms of open governance. The "Markets/contracts" form is a centrally controlled system where knowledge transfers are oriented at completed solutions, and open cooperation is regulated by clear contracts. The "partnerships/alliances" type allows for more open knowledge exchange and many communication channels in a diverse network of cooperators. Finally, the "user community-based" open governance is where the community generates solutions and manages the initiative. According to Felin and Zenger (2014), one of the central limitations of the purely community-based governance of innovation is the limited control over the development. This is because the selection of problems and solutions depends on individual users.

Which type of governance is best suited to open-source initiatives is an open debate. Most likely, it depends on the application and stage of the project. Central, top-down coordination can benefit open-source initiatives at various stages, but it should not dominate the management structure as it could block all bottom-up, non-deterministic open innovations. Eventually, the goal of open-source virtual collaborations is to create an ecosystem of capabilities (Abbate et al. 2021), where sharing internal and external knowledge allows to co-create novel solutions that exceed the potential of particular actors. Finding the balance in the strength and amount of coordination seems crucial for the success of open-source initiatives, especially in novel applications in social science. Given the limited evidence, we can only expect that more coordination is needed for the early stages of development, and it should be reduced (but not eliminated) once the bottom-up processes are activated.

Case study: the story of CPF survey harmonization

The following section will focus on the Comparative Panel File (CPF) as a case study of the first fully open-source survey harmonisation project in social science. I will present the origins and development of the CPF to illustrate the reasons, potentials and challenges of such an initiative. In particular, I will consider the role and challenge of coordination

Conventional approaches to survey harmonisation

Survey harmonisation is a process that creates a single data file out of various surveys that were mostly not designed to be integrated. The core task is to identify source variables referring to the same theoretical concept and transform them (e.g., recode and rescale) into target variables that can be compared across surveys (Wysmułek, Tomescu-Dubrow & Kwak 2021). This time-demanding and complex process involves technical (e.g., different data and file structures) and conceptual (e.g., different questionnaires and sample designs) challenges. Thus, the second-order knowledge infrastructure provided by data harmonisation can greatly benefit a broad research community. It saves weeks or months of harmonisation work, increases the usability of existing data, provides larger data sets, and stimulates new (e.g., comparative) applications. Over the past four decades, we observed various large-scale initiatives aimed at building a stable framework for *ex post* survey data harmonisation (Doiron et al. 2012; Dubrow & Tomescu-Dubrow 2016; Wolf et al. 2016; Wysmułek et al. 2021). For example, Survey Data Recycling (SDR) project (which builds on previous similar initiatives) creates a multi-country multi-year database pooled from cross-sectional surveys from over 150 countries and territories (Slomczynski & Tomescu-Dubrow 2018).

Ex-post data harmonization is particularly valuable for national panel studies that follow individuals or households over time. Many excellent and long-running panel studies around the world provide stable measurement over extensive periods (even decades), large samples and high response rates. Given the high costs of such surveys, their harmonisation can stimulate the cost-effective reuse of data, facilitate comparative research and contribute to the knowledge infrastructure of social science. However, there are still only a few *ex post* harmonisation initiatives of panel surveys in social science (Dubrow & Tomescu-Dubrow 2016). Despite differences in scale and scope, they all adopted the institutionalised and centralised organisational model, with a core development team separated from users, primarily top-down decision-making processes, and strongly embedded in public institutional frameworks (e.g. academic or government). As such, they are all prone to the limitations of the conventional model of knowledge infrastructure, namely, the path-dependent and closed developmental frameworks may limit their responsiveness and explanatory power.

Origins of the CPF

Together with Matthijs Kalmijn and Thomas Leopold, we initiated CPF in response to such limitations of conventional harmonisation initiatives (Turek et al. 2021). The story begins with a research idea about employment changes from a longitudinal and comparative perspective. A well-established and long-standing survey harmonisation project, the Cross-National Equivalent File (CNEF), appeared to us as a suitable data source (Burkhauser et al. 2001; Frick et al. 2007). CNEF integrates large household panel surveys from several countries in cooperation with the national source data administrators. It provides data files with harmonised outcome variables. The great comparative research value of the CNEF data has been confirmed in numerous publications, e.g., on income-related topics, life satisfaction, and self-employment (see: Turek et al., 2021).

However, it turned out that CNEF had limited applicability to the intended research. The first problem was the long waiting time for the data to arrive (several months). It can be a crucial barrier limiting scientific responsiveness for many applications. Furthermore, the latest panel waves of the original data were not yet integrated for some countries. Perhaps, the problem was related to the large organisational size; with several teams working worldwide, and complex, centralised management, CNEF's data production process may require more time. Like most scientific programs, CNEF also depends on governmental funding, which can affect the speed and direction of its further development. Yet the main disappointment was that the employment status – the central variable of interest – was not useful for the intended research due to how it was harmonised (e.g., unemployment was combined with non-employment). A necessary correction of the harmonization algorithm would be easy, but it was not possible in the CNEF system.

Although receiving ready-to-analysis data files is attractive, it also hampers harmonisation flexibility. Researchers cannot add or modify variables, even if the necessary information is available in the original datasets. Since the code is available only in parts and for selected countries, CNEF does not support modifications in harmonised variables or adding new variables from the source database directly by the users. Instead, CNEF's developers are responsible for providing credible and ultimate harmonisation solutions. However, multiple standards and strategies for *ex-post* harmonisation methodology exist (Dubrow & Tomescu-Dubrow 2016; Kołczyńska 2022). Household panel survey harmonization initiatives were also criticized for being dominated by a single field, i.e., economics, and focused on factors and variables relevant to economic research, such as income and wealth (Dubrow & Tomescu-Dubrow 2016). It is also a case of the CNEF that provides very detailed income-related indicators but focuses less on variables such as education, well-being, family relationship, or labour market status. As a result, despite the unique quality and potential, CNEF data have been applied to a relatively narrow set of research problems. Given the multitude of potential variables and research perspectives, centralized harmonisation will not fit all applications. It can also negatively affect transparency and learning processes that could help build common harmonisation standards (Wysmułek et al. 2021). As with the story of inequality research, conventional knowledge infrastructure may be efficient for certain research approaches, but path dependency and narrow perspectives will suppress other applications and constrain more general knowledge advancement.

We should admit that CNEF's authors have recognised these problems. They emphasised the need to keep the harmonisation process open, driven by the needs of research community, and possibly far from bureaucracy (Burkhauser & Lillard 2005; Frick et al. 2007). Burkhauser and Lillard (2005) refer to an example of the European Community Household Panel (ECHP), a prominent European harmonisation project that terminated prematurely in 2001. They argue that the failure was largely due to inefficient administration that ignored the research community's needs and could not integrate the project with a broader scientific infrastructure. CNEF has learned from these lessons and evolved over the years, e.g. extending explanatory power (by adding countries and variables) and aiming to implement more open communication. However, the entire content is still prepared by the CNEF team (even if inspired by researchers' needs), the main focus remains on incomes and earnings, and the project has not yet developed a fully bottom-up approach (Dubrow & Tomescu-Dubrow 2016). Thus, although CNEF is an extraordinary contribution to social science infrastructure, it cannot meet all expectations.

CPF's open-source model

This experience inspired my colleagues and me to reconsider the harmonisation model. CPF was an attempt to move the harmonisation process to open science and crowdsource cooperation and provide novel functionalities. CPF (www.cpfdata.com) is organised as a virtual platform that integrates tools for communication, code development, and general management of scientific research (for details, see Turek et al., 2021). The core of CPF is the freely and openly available harmonisation code built from scratch in Stata (one of the most popular statistical software). The code generates a comparative dataset based on the original household panel surveys (that are available for free from national data providers). The procedures integrate datasets and waves within countries, transform input variables into harmonised variables, and merge them into a single dataset. CPF version 1.4 data file contains ca. 2.8 million observations, coming from 370 thousand individuals and covering up to 41 waves per individual. Compared to CNEF (at least currently), CPF offers a different range of variables and more recent samples. The open-source code is organised into multiple lower- and higher-level files. It is stored at *GitHub*, a popular open-source code repository that provides tools to develop the code, track and share changes, and integrate them into consecutive versions. Users can modify and add variables, include more recent samples, or add new surveys.

Although CPF shares the same goal as its predecessor and focuses on the same datasets, the novel open-source framework and tools may contribute to comparative social science in several ways. At the basic level, CPF's open-source code can save weeks or months of harmonisation work. All household panel studies included in the CPF are extensively used in research and the comparative potential added by the CPF may only extend the utility of these surveys. Importantly, the code is also helpful for working with data from one country only.

From a broader perspective, CPF offers a novel open-source model for data infrastructure that stimulates new applications and extends researchers' flexibility. Compared to top-down initiatives, CPF allows for more open management, unconstrained development and better responsiveness to researchers' needs. The open-source format allows engaging the crowd wisdom (Beck et al. 2022) to boost creativity and extend the CPF code for new applications. The modularity of the process, i.e., decomposition of a complex harmonisation into more manageable and independent tasks (e.g., adding new variables), allows division of labour and parallel work. Once the coding framework that organises the most technical and time-consuming aspects of harmonisation (such as preparing and combining the source data files) is provided, researchers can focus on lower-scale tasks. Most of the distributed coding tasks in harmonisation can be classified as low complexity and well-structured, according to the terminology of Franzoni and Sauermann (2014). This means that tasks tend to be independent, and contributors can work in parallel. Such tasks refer to the most important input for the CPF, i.e., adding new variables and developing small parts of the code. They are organised by microtask workflows that instruct how to proceed. However, CPF's development can also involve highly complex and ill-structured tasks, such as adding new surveys or changing larger structures in the code. In this case, obtaining the final solution requires developing a common understanding of the goal and approach, sequential cooperation and coordinated verification of the changes.

Furthermore, the open and dynamic design supports technical and substantive solutions for harmonisation dilemmas. CNEF and other conventional harmonization projects intensively and successfully develop the harmonization methodology, focusing on providing unified, reliable and ready-to-analysis integrated datasets (e.g, Survey Data Recycling, see: Slomczynski & Tomescu-Dubrow 2018). Instead of top-down and ultimate solutions to comparability, CPF's open-source model explores a very different approach where researchers have complete control over the harmonisation process, yet they are also responsible for the quality and outcomes.

Inappropriate and erroneous harmonization is one of the major risks of the open-source model, yet transparency of the code facilitates error detection and code improvement. The coordination team can play an important role in these processes, however, the basic correcting mechanisms are crowd-based, assuming they are more efficient in this task (Arza et al. 2018). GitHub allows integration of the distributed code into consecutive official versions, providing version control. CPF is also integrated with Open Science Framework, one of the most popular open science platforms that facilitate collaborative workflow on research projects, pre-registering studies, storing code and data, etc. With permanent identifiers and continuous access to all versions of the data and documentation, the design stimulates transparency and reproducibility of research.

As a crowd-based cooperation, CPF can also be independent from the administrative and institutional constraints of conventional projects. The model can potentially improve the efficiency of harmonisation projects, and lower costs and time of comparative research. The cost of the CPF was incomparably lower than the cost of most data harmonisation initiatives, while providing comparable results. For example, the cost of building the *Consortium of Household Panels for European Socio-economic Research* (CHER) between 2000 and 2003 exceeded one million Euros (Dubrow & Tomescu-Dubrow 2016). As a comparison, the first published version of the CPF costs about 20 times less. Similar advantages are recognised in the case of virtual labs that highly decrease development costs and time, resulting in lower investment risks (Almaatouq et al. 2021). However, open-source projects are much more uncertain than institutionalized initiatives, so concerns about long-term sustainability remain a significant risk factor.

The CPF experiment so far

Open-source initiatives, such as CPF, appear and diffuse because such solutions are needed. But they will develop and mature only if they are useful for scientific research. CPF has already been recognised as a contribution to the research infrastructure in social science. It received a positive response from data providers, scholars and research institutions. In the first ca. two years after the publication in December 2020, the interest was substantial, with ca. 10,000+ site views from 100+ countries, 30,000+ social media interactions, and 6,400 views of the main article (Turek et al., 2021). Nevertheless, CPF is still in its initial stage, so assessing its contribution to social science and reflecting on the open-source model as an alternative to conventional institutionalised solutions is difficult. The main evaluation criteria should be twofold.

First, the major scientific criterion is related to applications of the infrastructure by researchers and can be measured by scientific output, e.g., publications. Previous harmonization initiatives (e.g., CNEF, LIS, and ECHP) were very successful in this regard, leading to many publications. As the publication process is lengthy in science, two-three years is too early to evaluate it, yet several articles have already been published, e.g., (Thielemans & Mortelmans 2022; Turek, Henkens & Kalmijn 2022). Central for good publications will be the usefulness of the code, quality of the harmonization, and researchers' flexibility.

The second criterion considers the active involvement of the academic community in code development. So far, the community-based input has been limited but not negligible. CPF received several substantial contributions from active users (e.g., error detection) and external developers (e.g., pieces of code, detailed suggestions). We also initiated larger cooperations aimed at extensive developments. However, it is too early to judge whether the CPF's open-source model is successful in this dimension. A central question remains of how much coordination is required for the project to grow. Until now, the CPF team has led most of the ongoing development[1]. Stimulating and integrating external developers at a larger scale in developing processes is fundamental to success (Duparc et al. 2022; Setia et al. 2012). One concern is the usability of the GitHub environment for crowd-based code development, which seems challenging for many users who prefer to share ideas by email rather than introduce them directly in the code. In particular, more demanding tasks (like adding new countries) appear too complex for purely crowd-based cooperation. Alternative solutions for active development and technical improvements of website usability should be considered in the future.

Conclusions

Open-source collaborations are still rarely encountered in social research, however they present an attractive organisational scheme. Given the rapid social changes and growing complexity of knowledge production, systemic constraints remain to the quality, innovativeness, efficiency and speed of knowledge production in social science. The history of inequality studies revealed the limitations and malfunctions of the conventional, institutionalised academia, such as narrow approaches, path-dependent development, and slow reaction to social problems.

The article discussed the coordinated open-source model as an attempt to open the knowledge infrastructure. Flexibility, decentralised control, and community-based development facilitate breaking down path dependencies, opening possibilities for innovations and helping respond to societal challenges. As a result, the model can contribute to the explanatory power and responsiveness of social science. It also aligns with the ideas of open science, such as reproducibility,

[1] E.g., CPF v.1.5 (April 2023) has been prepared by I. Voets under supervision of K. Turek and M. Kalmijn at NIDI, partly based on users suggestions and input.

transparency and accessibility, which can stimulate more robust and impactful findings and be appealing to funding agencies. However, the lessons learned from non-scientific open-source initiatives, e.g., software programming, point to various risk factors that often lead to project abandonment. Thus, I argued that open-source initiatives require a certain amount of coordination to provide stability and stimulate development. The coordinated open-source model can be potentially applied in certain areas of the increasingly computational and coding-dependent social science, serving as an important addition to conventional research infrastructure. Although, it is too early to evaluate their impact, it is worth observing such initiatives.

Bibliography

Abbate, T., Codini, A., Aquilani, B., and Vrontis, D. (2021). From Knowledge Ecosystems to Capabilities Ecosystems: When Open Innovation Digital Platforms Lead to Value Co-creation. *Journal of the Knowledge Economy* 13(1): 290-304.

Acker, J. (2016). Inequality Regimes: Gender, Class, and Race in Organizations. *Gender & Society* 20(4): 441-64.

Alipourfard, N., Arendt, B., Benjamin, D.M., Benkler, N., Bishop, M.M., Burstein, M., . . . Wu, J. (2021). Systematizing Confidence in Open Research and Evidence (SCORE). *SocArXiv. May 4. doi:10.31235/osf.io/46mnb.*

Almaatouq, A., Becker, J.A., Bernstein, M.S., Botto, R., Bradlow, E., Damer, E., . . . Yin, M. (2021). Scaling up experimental social, behavioral, and economic science. *Preprint: OSF. https://doi.org/10.17605/OSF.IO/KNVJS*.

Altman, M., and Cohen, P.N. (2022). The Scholarly Knowledge Ecosystem: Challenges and Opportunities for the Field of Information. *Frontiers in Research Metrics Analytics* 6: 751553.

Amis, J.M., Mair, J., and Munir, K.A. (2020). The Organizational Reproduction of Inequality. *Academy of Management Annals* 14(1): 195-230.

Anderson, P. (1999). Perspective: Complexity Theory and Organization Science. *Organization Science* 10(3): 216-32.

Anderson, P.W. (1972). More Is Different: Broken symmetry and the nature of the hierarchical structure of science. *Science* 177(4047): 393-96.

Arthur, W.B. (1994). *Increasing Returns and Path Dependence in the Economy*. Ann Arbor: University of Michigan Press.

Arza, V., Fressoli, M., Chan, L., and Loizides, F. (2018). Systematizing benefits of open science practices. *Information Services & Use* 37(4): 463-74.

Auspurg, K., and Brüderl, J. (2021). Has the Credibility of the Social Sciences Been Credibly Destroyed? Reanalyzing the "Many Analysts, One Data Set" Project. *Socius* 7: 1-14.

Avelino, G.A., Constantinou, E., Valente, M.T., and Serebrenik, A. (2019). On the abandonment and survival of open source projects: an empirical investigation. *Proceedings - 13th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, IEEE Computer Society*.

Aydinoglu, A.U. (2013). Toward a New Understanding of Virtual Research Collaborations. *SAGE Open* 3(4): 1-12.

Bak-Coleman, J.B., Alfano, M., Barfuss, W., Bergstrom, C.T., Centeno, M.A., Couzin, I.D., . . . Weber, E.U. (2021). Stewardship of global collective behavior. *PNAS* 118(27): 1-10.

Baldwin, C.Y., and Clark, K.B. (2006). The Architecture of Participation: Does Code Architecture Mitigate Free Riding in the Open Source Development Model? *Management Science* 52(7): 1116-27.

Banal-Estañol, A., Macho-Stadler, I., and Pérez-Castrillo, D. (2019). Evaluation in research funding agencies: Are structurally diverse teams biased against? *Research Policy* 48(7): 1823-40.

Baron, J., and Bielby, W. (1980). Bringing the firms back in -stratification, segmentation, and the organization. *American Sociological Review* 45(5): 737-65.

Beck, S., Brasseur, T.-M., Poetz, M., and Sauermann, H. (2022). Crowdsourcing research questions in science. *Research Policy* 51(4).

Benbya, H., Jacucci, E., and McKelvey, B. (2006). Toward a complexity theory of information systems development. *Information Technology & People* 19(1): 12-34.

Besancon, L., Peiffer-Smadja, N., Segalas, C., Jiang, H., Masuzzo, P., Smout, C., . . . Leyrat, C. (2021). Open science saves lives: lessons from the COVID-19 pandemic. *BMC Medical Research Methodology* 21(117): 1-18.

Blundell, R., Costa Dias, M., Cribb, J., Joyce, R., Waters, T., Wernham, T., and Xu, X. (2022). Inequality and the COVID-19 Crisis in the United Kingdom. *Annual Review of Economics* 14(1): 607-36.

Bonaccorsi, A., and Rossi, C. (2003). Why Open Source software can succeed. *Research Policy* 32(7): 1243-58.

Breznau, N. (2021). Does Sociology Need Open Science? *Societies* 11(1): 1-25.

Brüderl, J., Kratz, F., and Bauer, G. (2019). Life course research with panel data: An analysis of the reproduction of social inequality. *Advances in Life Course Research* 41.

Burkhauser, R.V., Butrica, B.A., Daly, M.C., and Lillard, D.R. (2001). The Cross-National Equivalent File: A product of cross-national research. in *Social Insurance in a Dynamic Society*, edited by Becker, I., Ott, N., and Rolf, G. Frankfurt: Campus Fachbuch.

Burkhauser, R.V., and Lillard, D.R. (2005). The contribution and potential of data harmonization for cross-national comparative research. *Journal of Comparative Policy Analysis: Research and Practice* 7(4): 313-30.

Callaway, E. (2020). Will the pandemic permanently alter scientific publishing? *Nature* 582: 167-68.

Chen, X., Li, X., Clark, J.G., and Dietrich, G.B. (2013). Knowledge sharing in open source software project teams: A transactive memory system perspective. *International Journal of Information Management* 33(3): 553-63.

Chesbrough, H. (2003). *Open Innovation: The New Imperative for Creating and Profiting from Technology*: Harvard Business Press.

Cohen, W., and Levinthal, D. (1990). Absorptive capacity - a new perspective on learning and innovation. *Administrative Science Quarterly* 30(1): 128-52.

Connelly, R., Playford, C.J., Gayle, V., and Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research* 59: 1-12.

Coombs, R., and Hull, R. (1998). 'Knowledge management practices' and path-dependency in innovation. *Research Policy* 28: 237–53.

Crossley, N. (2021). A Dependent Structure of Interdependence: Structure and Agency in Relational Perspective. *Sociology* 56(1): 166-82.

Crowston, K., Wei, K., Howison, J., and Wiggins, A. (2012). Free/Libre open-source software development. *ACM Computing Surveys* 44(2): 1-35.

Crystal, S., and Shea, D. (1990). Cumulative advantage, cumulative disadvantage, and inequality among elderly people. *The Gerontologist* 30(4): 437-43.

Dahlander, L., and Gann, D.M. (2010). How open is innovation? *Research Policy* 39(6): 699-709.

Dahlander, L., Gann, D.M., and Wallin, M.W. (2021). How open is innovation? A retrospective and ideas forward. *Research Policy* 50(4).

David, P.A. (2007). Path dependence: a foundational concept for historical social science. *Cliometrica* 1(2): 91-114.

Devillers, J., Devillers, H., Decourtye, A., and Aupinel, P. (2010). Internet resources for agent-based modelling. *SAR QSAR Environ Res* 21(3-4): 337-50.

DiPrete, T.A., and Fox-Williams, B.N. (2021). The Relevance of Inequality Research in Sociology for Inequality Reduction. *Socius* 7: 1-31.

Doiron, D., Raina, P., Raina, P., L'Heureux, F., and Fortier, I. (2012). Facilitating collaborative research: Implementing a platform supporting data harmonization and pooling. *Norsk Epidemiologi* 21(2): 221-24.

Dubrow, J.K., and Tomescu-Dubrow, I. (2016). The rise of cross-national survey data harmonization in the social sciences: emergence of an interdisciplinary methodological field. *Quality & Quantity* 50(4): 1449-67.

Duparc, E., Möller, F., Jussen, I., Stachon, M., Algac, S., and Otto, B. (2022). Archetypes of open-source business models. *Electronic Markets* 32(2): 727-45.

Edelmann, A., Wolff, T., Montagne, D., and Bail, C.A. (2020). Computational Social Science and Sociology. *Annual Review of Sociology* 46(1): 61-81.

Edwards, P.N., Jackson, S.J., Chalmers, M.K., Bowker, G.C., Borgman, C.L., Ribes, D., . . . Calvert, S. (2013). *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges*. Ann Arbor: Deep Blue.

Elder-Vass, D. (2010). *The causal power of social structures : emergence, structure and agency*. Cambridge ; New York: Cambridge University Press.

Esposito De Falco, S., Renzi, A., Orlando, B., and Cucari, N. (2017). Open collaborative innovation and digital platforms. *Production Planning & Control* 28(16): 1344-53.

EU. (2016). *Realising the European open science cloud*: European Commission, Directorate-General for Research and Innovation.

Evans, J., and Foster, J. (2011). Metaknowledge. *Science* 331: 721-25.

Faraj, S., and Sproull, L. (2000). Coordinating Expertise in Software Development Teams. *Management Science* 46(12): 1554-68.

Fecher, B., Friesike, S., and Hebing, M. (2015). What drives academic data sharing? *PLoS One* 10(2): e0118053.

Fecher, B., and Wagner, G. (2016). Open Access, Innovation, and Research Infrastructure. *Publications* 4(2): 17.

Felin, T., and Zenger, T.R. (2014). Closed or open innovation? Problem solving and the governance choice. *Research Policy* 43(5): 914-25.

Firebaugh, G. (2007). Replication Data Sets and Favored-Hypothesis Bias. *Sociological Methods & Research* 36(2): 200-09.

Franck, R.E. (2002). *The Explanatory Power of Models: Bridging the Gap between Empirical and Theoretical Research in the Social Sciences*. New York: Springer Science & Business Media.

Franzoni, C., Poetz, M., and Sauermann, H. (2021). Crowds, citizens, and science: a multi-dimensional framework and agenda for future research. *Industry and Innovation* 29(2): 251-84.

Franzoni, C., and Sauermann, H. (2014). Crowd science: The organization of scientific research in open collaborative projects. *Research Policy* 43(1): 1-20.

Fraser, N., Brierley, L., Dey, G., Polka, J., Pálfy, M., Nanni, F., and Coates, J. (2021). The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLoS Biol* 19(4): e3000959.

Freese, J. (2007). Replication standards for quantitative social science: Why not sociology? *Sociological Methods & Research* 36(2): 153-72.

Freese, J., and King, M.M. (2018). Institutionalizing Transparency. *Socius: Sociological Research for a Dynamic World* 4: 237802311773921.

Freese, J., Rauf, T., and Voelkel, J.G. (2022). Advances in transparency and reproducibility in the social sciences. *Social Science Research* 107: 102770.

Frick, J., Jenkings, S.P., Lillard, D.R., Lipps, O., and Wooden, M. (2007). The Cross-National Equivalent File (CNEF) and Its Member Country Household Panel Studies. *EconStor Open Access Articles, ZBW - Leibniz Information Centre for Economics*: 627-54.

Friesike, S., Widenmayer, B., Gassmann, O., and Schildhauer, T. (2014). Opening science: towards an agenda of open science in academia and industry. *The Journal of Technology Transfer* 40(4): 581-601.

Galea, S., Riddle, M., and Kaplan, G.A. (2010). Causal thinking and complex system approaches in epidemiology. *International Journal of Epidemiology* 39(1): 97-106.

Gangl, M. (2010). Causal Inference in Sociological Research. *Annual Review of Sociology* 36(1): 21-47.

Gerring, J., Mahoney, J., and Elman, C. (2020). *The Production of Knowledge: Enhancing Progress in Social Science*. Cambridge: Cambridge University Press.

Habermas, J. (1984). *Theory of Communicative Action: Reason and the Rationalization of Society*. Boston, MA: Becon Press.

Hirschman, D. (2021). Rediscovering the 1%: Knowledge Infrastructures and the Stylized Facts of Inequality. *American Journal of Sociology* 127(3): 739-86.

Hofman, J.M., Sharma, A., and Watts, D.J. (2017). Prediction and explanation in social systems. *Science* 355: 486–88.

Hofman, J.M., Watts, D.J., Athey, S., Garip, F., Griffiths, T.L., Kleinberg, J., . . . Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature* 595(7866): 181-88.

Hollingsworth, J.R. (2008). Scientific Discoveries: An Institutionalist and Path-Dependent Perspective. Pp. 317-53, edited by Hannaway, C. Amsterdam: IOS Press.

Horton, J.J., Rand, D.G., and Zeckhauser, R.J. (2011). The online laboratory: conducting experiments in a real labor market. *Experimental Economics* 14(3): 399-425.

Hucka, M., Nickerson, D.P., Bader, G.D., Bergmann, F.T., Cooper, J., Demir, E., . . . Le Novere, N. (2015). Promoting Coordinated Development of Community-Based Information Standards for Modeling in Biology: The COMBINE Initiative. *Front Bioeng Biotechnol* 3: 19.

Jackson, M. (2022). How Is It To Be Done? Building a Social Science of Radical Reform. *Socius* 8: 1-6.

Jain, U., Min, J., and Lee, J. (2016). Harmonization of cross-national studies of aging to the Health and Retirement Study - user guide: Family transfer - informal care. *University of Southern California, CESR-Schaeffer Working Paper Series No. 2016-008*.

Janssen, M.A., Alessa, L.N.I., Barton, M., Bergin, S., and Lee, A. (2008). Towards a Community Framework for Agent-Based Modelling. *Journal of Artificial Societies and Social Simulation* 11(2).

Jones, B. (2009). The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder? *The Review of Economic Studies* 76(1): 283-317.

Kapiszewski, D., and Karcher, S. (2020). Making Research Data Accessible. Pp. 197-220 in *The Production of Knowledge: Enhancing Progress in Social Science*, edited by Elman, C., Gerring, J., and Mahoney, J. Cambridge Cambridge University Press.

King, G. (1995). Replication, Replication. *PS: Political Science and Politics* 28(3): 444-52.

—. (2007). An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. *Sociological Methods & Research* 36(2): 173-99.

—. (2011a). Ensuring the Data-Rich Future of the Social Sciences. *Science* 331(11): 719-21.

—. (2011b). Ensuring the data-rich future of the social sciences. *Science* 331: 719-21.

Kołczyńska, M. (2022). Combining multiple survey sources: A reproducible workflow and toolbox for survey data harmonization. *Methodological Innovations* 15(1): 62-72.

Krücken, G. (2003). Learning the 'New, New Thing': On the role of path dependency in university structures. *Higher Education* 46: 315–39.

Kühne, S., Kroh, M., Liebig, S., and Zinn, S. (2020). The Need for Household Panel Surveys in Times of Crisis: The Case of SOEP-CoV. *Survey Research Methods* 14(2).

Kullenberg, C., and Kasperowski, D. (2016). What Is Citizen Science? A Scientometric Meta-Analysis. *PLoS One* 11(1): e0147152.

Lazer, D., Pentland, A., Watts, D., Aral, S., Athey, S., Contractor, N., . . . Margetts, H. (2020). Computational social science: obstacles and opportunities. *Science* 369(6507): 1060–62.

Lehtola, S., and Karttunen, A.J. (2022). Free and open source software for computational chemistry education. *WIREs Computational Molecular Science* 12(5): 1-33.

Linåker, J., and Regnell, B. (2020). What to share, when, and where: balancing the objectives and complexities of open source software contributions. *Empirical Software Engineering* 25(5): 3799-840.

Low-Décarie, E., Chivers, C., and Granados, M. (2014). Rising complexity and falling explanatory power in ecology. *Frontiers in Ecology and the Environment* 12(7): 412-18.

Lucas-Dominguez, R., Alonso-Arroyo, A., Vidal-Infer, A., and Aleixandre-Benavent, R. (2021). The sharing of research data facing the COVID-19 pandemic. *Scientometrics* 126(6): 4975-90.

Macy, M.W., and Willer, R. (2002). From Factors to Actors: Computational Sociology and Agent-Based Modeling. *Annual Review of Sociology* 28(1): 143-66.

Marwick, B. (2016). Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation. *Journal of Archaeological Method and Theory* 24(2): 424-50.

Mason, W., and Watts, D.J. (2012). Collaborative learning in networks. *Proc Natl Acad Sci U S A* 109(3): 764-9.

Matei, A., and Irimia, S.I. (2014). Open Source Governance—A More Ambitious Cousin of Collaborative Governance. *International Journal of Public Administration* 37(12): 812-23.

Merton, R.K. (1973 [1942]). The Normative Structure of Science. Pp. 267–78 in *The Sociology of Science: Theoretical and Empirical Investigations*, edited by Merton, R.K. and Storer, N.W. Chicago University of Chicago Press.

Moshontz, H., Campbell, L., Ebersole, C.R., H, I.J., Urry, H.L., Forscher, P.S., . . . Chartier, C.R. (2018). The Psychological Science Accelerator: Advancing Psychology through a Distributed Collaborative Network. *Advances in Methods and Practices in Psychological Science* 1(4): 501-15.

Nosek, B.A., Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D., Breckler, S.J., . . . Yarkoni, T. (2015). Promoting an open research culture. *Science* 348(6242): 1422-25.

Olson, M. (1965). *The Logic of Collective Action*. Cambridge, MA: Harvard University Press.

Peirce, C.S. (1902). Truth and Falsity and Error. *Dictionary of Philosophy and Psychology*: 718-20.

Pfaff, C., and Hasan, H. (2007). Can Knowledge Management be Open Source. Pp. 59-70 in *The International Federation for Information Processing*, edited by Feller, J.e.a. Boston: Springer.

Piccarreta, R., and Studer, M. (2019). Holistic analysis of the life course: Methodological challenges and new perspectives. *Advances in Life Course Research*.

Piketty, T. (2014). *Capital in the Twenty-First Century*. Cambridge Massachusetts: Belknap Press.

Piketty, T., and Saez, E. (2003). Income Inequality in the United States, 1913–1998. *The Quarterly Journal of Economics* 18(1): 1–41.

Popper, K.R. (1959 [1934]). *The Logic of Scientific Discovery*. New York: Basic Books.

Rai, A. (2005). *Open and Collaborative Research - A New Model for Biomedicine.pdf*>.

Ray, V. (2019). A Theory of Racialized Organizations. *American Sociological Review* 84(1): 26-53.

Riaz, S. (2015). Bringing inequality back in: The economic inequality footprint of management and organizational practices. *Human Relations* 68(7): 1085-97.

Salganik, M. (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.

Salganik, M.J., Dodds, P.S., and Watts, D.J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311(5762): 854-6.

Salganik, M.J., and Levy, K.E. (2015). Wiki surveys: open and quantifiable social data collection. *PLoS One* 10(5): e0123483.

Salganik, M.J., Lundberg, I., Kindel, A.T., Ahearn, C.E., Al-Ghoneim, K., Almaatouq, A., . . . McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proc Natl Acad Sci U S A* 117(15): 8398-403.

Savage, M. (2021). The return of inequality. in *The Return of Inequality*. Harvard Harvard University Press.

Scheliga, K., Friesike, S., Puschmann, C., and Fecher, B. (2018). Setting up crowd science projects. *Public Understanding of Science* 27(5): 515-34.

Sen, A. (1993). Capability and Well-Being. Pp. 30–53 in *The Quality of Life* edited by Nussbaum, M. and Sen, A. Oxford: Oxford Academic.

Setia, P., Rajagopalan, B., Sambamurthy, V., and Calantone, R. (2012). How Peripheral Developers Contribute to Open-Source Software Development. *Information Systems Research* 23(1): 144-63.

Shah, S.K. (2006). Motivation, Governance, and the Viability of Hybrid Forms in Open Source Software Development. *Management Science* 52(7): 1000-14.

Slomczynski, K., and Tomescu-Dubrow, I. (2018). Basic Principles of Survey Data Recycling. Pp. 937-62 in *Advances in Comparative Survey Methodology: Multinational, Multiregional and Multicultural Contexts* edited by T.P. Johnson, B.-E.P., I. A. L. Stoop, B. Dorer. New Jersey: Wiley Hoboken.

Starbuck, W.H. (2006). *The Production of Knowledge: The Challenge of Social Science*. Oxford: Oxford University Press.

Steinbacher, M., Raddant, M., Karimi, F., Camacho Cuena, E., Alfarano, S., Iori, G., and Lux, T. (2021). Advances in the agent-based modeling of economic and social behavior. *SN Bus Econ* 1(7): 99.

Steinmacher, I., Graciotto Silva, M.A., Gerosa, M.A., and Redmiles, D.F. (2015). A systematic literature review on the barriers faced by newcomers to open source software projects. *Information and Software Technology* 59: 67-85.

Sterman, J.D., and Wittenberg, J. (1999). Path Dependence, Competition, and Succession in the Dynamics of Scientific Revolution. *Organization Science* 10(3): 322-41.

Swedberg, R. (2020). Exploratory Research. Pp. 17-41 in *The Production of Knowledge: Enhancing Progress in Social Science*, edited by Elman, C., Gerring, J., and Mahoney, J. Cambridge Cambridge University Press.

Thielemans, G., and Mortelmans, D. (2022). Poverty Risks after Relationship Dissolution and the Role of Children: A Contemporary Longitudinal Analysis of Seven OECD Countries. *Social Sciences* 11(3).

Tomaskovic-Devey, D., and Avent-Holt, D. (2019). *Relational Inequalities: An organizational approach*. Oxford: Oxford University Press.

Turek, K., Henkens, K., and Kalmijn, M. (2022). Gender and Educational Inequalities in Extending Working Lives: Late-Life Employment Trajectories Across Three Decades in Seven Countries. *Work, Aging and Retirement* (waac021).

Turek, K., Kalmijn, M., and Leopold, T. (2021). The Comparative Panel File: Harmonized Household Panel Surveys from Seven Countries. *European Sociological Review* 37(3): 505-23.

Uhlmann, E.L., Ebersole, C.R., Chartier, C.R., Errington, T.M., Kidwell, M.C., Lai, C.K., . . . Nosek, B.A. (2019). Scientific Utopia III: Crowdsourcing Science. *Perspect Psychol Sci* 14(5): 711-33.

Valentine, M.A., Retelny, D., To, A., Rahmati, N., Doshi, T., and Bernstein, M.S. 2017. "Flash Organizations." Pp. 3523-37 in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*.

Van De Ven, A., and Johnson, P.E. (2006). Knowledge for Theory and Practice. *The Academy of Management Review* 31(4): 802-21.

Vazire, S. (2018). Implications of the Credibility Revolution for Productivity, Creativity, and Progress. *Perspect Psychol Sci* 13(4): 411-17.

Volberda, H., Schneidmuller, T., and Zadeh, T. (2021). Knowledge and Innovation - From Path Dependency Toward Managerial Agency. Pp. 445-66 in *Strategic Management: State of the Field and Its Future*, edited by Duhaime, I.M., Hitt, M.A., and Lyles, M.A. New York, NY: Oxford University Press.

Watson, C. (2022). Rise of the preprint: how rapid data sharing during COVID-19 has changed science forever. *Nature Medicine* 28(1): 2-5.

Wolf, C., Schneider, S., Behrand, D., and Joye, D. (2016). Harmonizing survey questions between cultures and over time. Pp. 502-24 in *The SAGE Handbook of Survey Methodology*, edited by Wolf, C., Joye, D., Smith, T., and Fu, Y.-c. Los Angeles SAGE.

Woolley, R., Sánchez-Barrioluengo, M., Turpin, T., and Marceau, J. (2015). Research collaboration in the social sciences: What factors are associated with disciplinary and interdisciplinary collaboration? *Science and Public Policy* 42(4): 567-82.

Wuchty, S., Jones, B.F., and Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. *Science* 316: 1036-39.

Wysmułek, I., Tomescu-Dubrow, I., and Kwak, J. (2021). Ex-post harmonization of cross-national survey data: advances in methodological and substantive inquiries. *Quality & Quantity* 56(3): 1701-08.

Zuo, Z., and Zhao, K. (2018). The more multidisciplinary the better? – The prevalence and interdisciplinarity of research collaborations in multidisciplinary institutions. *Journal of Informetrics* 12(3): 736-56.