EUROPEAN SOCIAL SCIENCE INFRASTRUCTURE

DRAFT WHITE PAPER

VERSION 0.1
9 DECEMBER 2023

EXECUTIVE SUMMARY

Building on the achievements of recent years and envisioning a stronger alignment between the needs of the research community, we offer 7 clear recommendations for developments in Data, Computation, FAIR, Tools, Expertise, and Governance.

Greater investments in European Social Science Infrastructure will make use of the opportunities of new data, methods and tools whilst also rising collectively to the challenges that they pose, harnessing the power of this ongoing data revolution.

The challenges we face, such as climate change, ethical AI, misinformation, and population ageing, all have an international character

This white paper has been prepared by Tom Emery, Kasia Karpinska, Angelica Maineri and Lucas van der Meer and is open to comments, contributions, and co-authorship. We are aiming to collate the infrastructural needs of the European Social Science community in a coherent plan.

If we have made errors, omissions, or misrepresented these needs the we would be very happy to correct this. All comments and contributions can be sent to tom@odissei-data.nl and will be very gratefully received.

INTRODUCTION

Many of the biggest questions faced by European society today are ones that are intrinsically social and cross-border. How can we develop an economy that is ecologically sustainable? How can we reduce inequalities and improve the opportunities of all members of society? How can we ensure that technological shifts in how we communicate bring us closer together and not push us further apart?

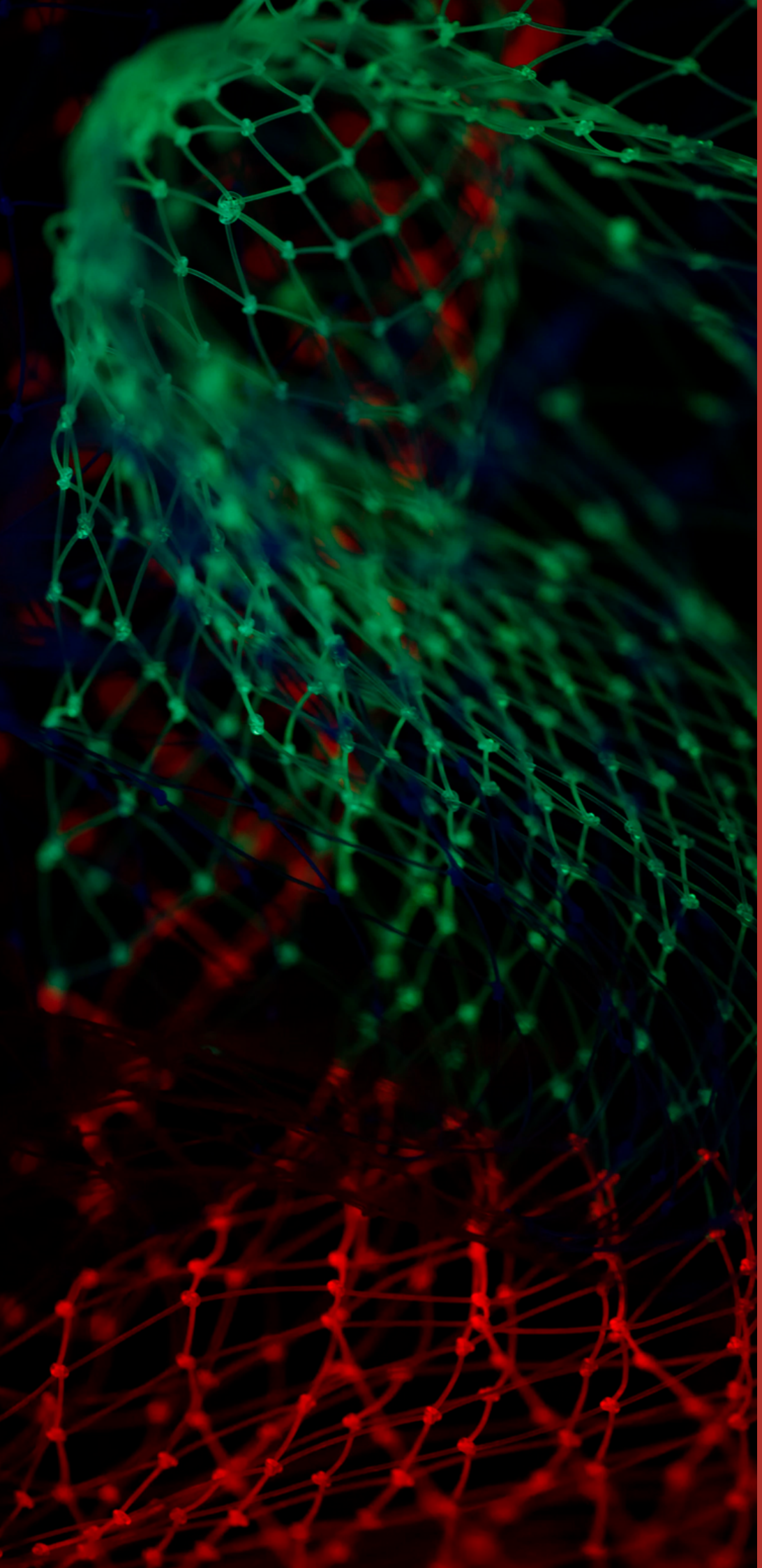
Because of these questions and many more, it is an exciting time to be a social scientist. New data, methods, tools and computational resources are opening up new lines of enquiry every day. But with these innovations come challenges on how to apply new methods rigorously, to judge new data forms ethically, and to ensure the sustainability of our tools and computational resources. Because of the immediacy of these opportunities and the scale of these challenges, it is more vital than ever that we see greater investment in European Social Science Infrastructure.

Social Science Infrastructure will allow us to make use of the opportunities of new data, methods, and tools whilst also rising collectively to the challenges that they pose. Scale and interoperability are needed to achieve this, and efforts that are concentrated at the national level or in specific domains will not be sufficient. These challenges stretch across European Society and permeate every domain of social science. We must invest in a better understanding of human society before it's too late.

To achieve this it is vital that we develop a strategy that reflects how much we have already achieved, how we can build on these successes, and how we can better align the investments in European Social Science Infrastructure with the needs of the scientific community. In this white paper we set out what such a strategy could look like and bring these together in *7 clear recommendations* for the future development of European Social Science Infrastructure.

These recommendations are guided by widely held commitments to open science as well as the FAIR and FACT principles. We also recognize that there are many exceptional elements in the European landscape, including the European Open Science Cloud and Social Science and Humanities Open Cluster. In drafting these recommendations, we seek to build on these successes, to enhance and complement their existing sizeable impact.

This white paper is not an endpoint. Instead, we hope it is the start of a broader discussion on what European Social Science needs in order to meet these grand societal challenges. We appreciate engagement and feedback of all forms.



DATA

PAGES
3-5

RECOMMENDATION #1: CREATE A PAN-EUROPEAN PANEL SURVEY

A pan-European probability-based online panel should be created that covers all countries, is conducted monthly, enables modularity, and includes the integration of other platforms for non-survey-based experiments and data donation.

The panel should be available to researchers through a) an open, excellence-based grant system, and b) a paid basis of access to generate revenues.

The panel would also provide access on a paid basis to the European Commission and national governments to supplement the functionality of Eurobarometer and offer higher scientific standards.

The panel would be designed in the same form as current national online panels such as the GESIS Panel, ELIPSS Panel, and LISS panel which have already demonstrated the high level of demand for such data collection facilities. For the offline population, specific support would be provided to enable their participation as is the case in the UKLHS, CRONOS, and many other online surveys.

The panel could also serve as the basis for or complement to the existing European Survey Infrastructures, including the European Social Survey, the European Values Study, and the Generations and Gender Programme. It would be based-off a centralized system operated via a single site, like the service provided by Centerdata in the implementation of the Survey of Health, Ageing, and Retirement in Europe.

Such an online panel would, for the first time, offer 'beam-time' to early career researchers looking to field their own survey items and experiments to a high-quality, pan-European panel. This would greatly democratize, accelerate innovation, and improve the openness of European Social Science.

All data collected in the panel would be open to verified researchers using a lightweight and simple access procedure.

Such a panel would revolutionize the field of Social Science and allow researchers to collect high-quality data on pertinent social issues rapidly. This was exemplified during the pandemic when such panels enabled researchers to understand life during lockdown and vital issues such as vaccine hesitancy.

A European Online Panel would allow Social Scientists to keep their finger on the pulse of Europe.

RECOMMENDATION #2: ESTABLISH A EUROPEAN INTERNET OBSERVATORY

Much of the data infrastructure in Europe at the moment is designed to support data directly collected by scientists such as survey data. However, the social sciences have been radically transformed over the past 20 years by the increasing availability of social media data and digital trace data for the study of human behaviour.

Despite this growth, the recent Twitter API shutdown has had dramatic consequences for existing research agendas and the social science community's dependence on these platforms collaborations has been exposed.

Europe poses a unique competitive advantage provided for through GDPR and the Data Governance Act with which to respond to this. European Data Legislation ensures that all individuals must be able to access a digital download package in a machine readable format which provides a transparent, legally grounded, and scalable way for data to be collected on individuals via 'data donation'.

The software, expertise and infrastructure needed to coordinate, harvest, and collate these data donation efforts are currently a vast unmet need of the social science community. The demands for this infrastructure are evident from infrastructures such as SoBigData Europe, GUIDE, and MeDem which seek to combine data from a wide variety of sources. However, these projects do not offer infrastructure for collecting and harmonizing this data directly.

This trend in the use of social media and digital trace data is reflected in significant large scale investments at the national level such as the Leibniz Associationsupport for the creation of the Digital Behavioural Data Pillar at GESIS and the ESRC investment in Smart Data Research UK. These groundbreaking national initiatives need to be better coordinated and integrated with existing infrastructures and to engage with the social media platforms that are the source of this data.

Just as cross-national surveys have been developed to examine social questions at the European level, it is vital for a digital Europe to have a European Level research facility. We refer to this facility as the European Internet Observatory in reference to the National Internet Observatory in the United States which uses a data donation framework to monitor the online behaviour of thousands of volunteers.

We propose that a similar facility be developed for Europe, and specifically one that seeks to integrate digital trace data across borders and align data sources so that research can focus on the European dimension to these questions in compliance with GDPR. A single facility is vital in this area as trust is vital to the concept of data donation, and visible, established, and clearly signposted facilities make this process more transparent and trustworthy. We would also strongly urge that such a facility be closely aligned with existing and future survey infrastructures so that digital trace data can be collected in these infrastructures, connected to high-quality survey data, and the broader impact of online life be better understood.

COMPUTATION

PAGES
6-7

RECOMMENDATION #3: INVEST IN AN INTERNATIONAL DATA ACCESS NETWORK

Social scientists regularly encounter computational constraints, especially when dealing with large scale and complex datasets. However, the primary challenge for social scientists is not in sourcing computational resources but ensuring that these computational resources comply with the security standards of existing national and European Legislation.

This has led to the wide scale development of 'Trusted Research Environments' (TREs). These environments allow researchers to conduct their analysis on a virtual machine which can be resourced with a very large amount of computational power as well as the required data and tools.

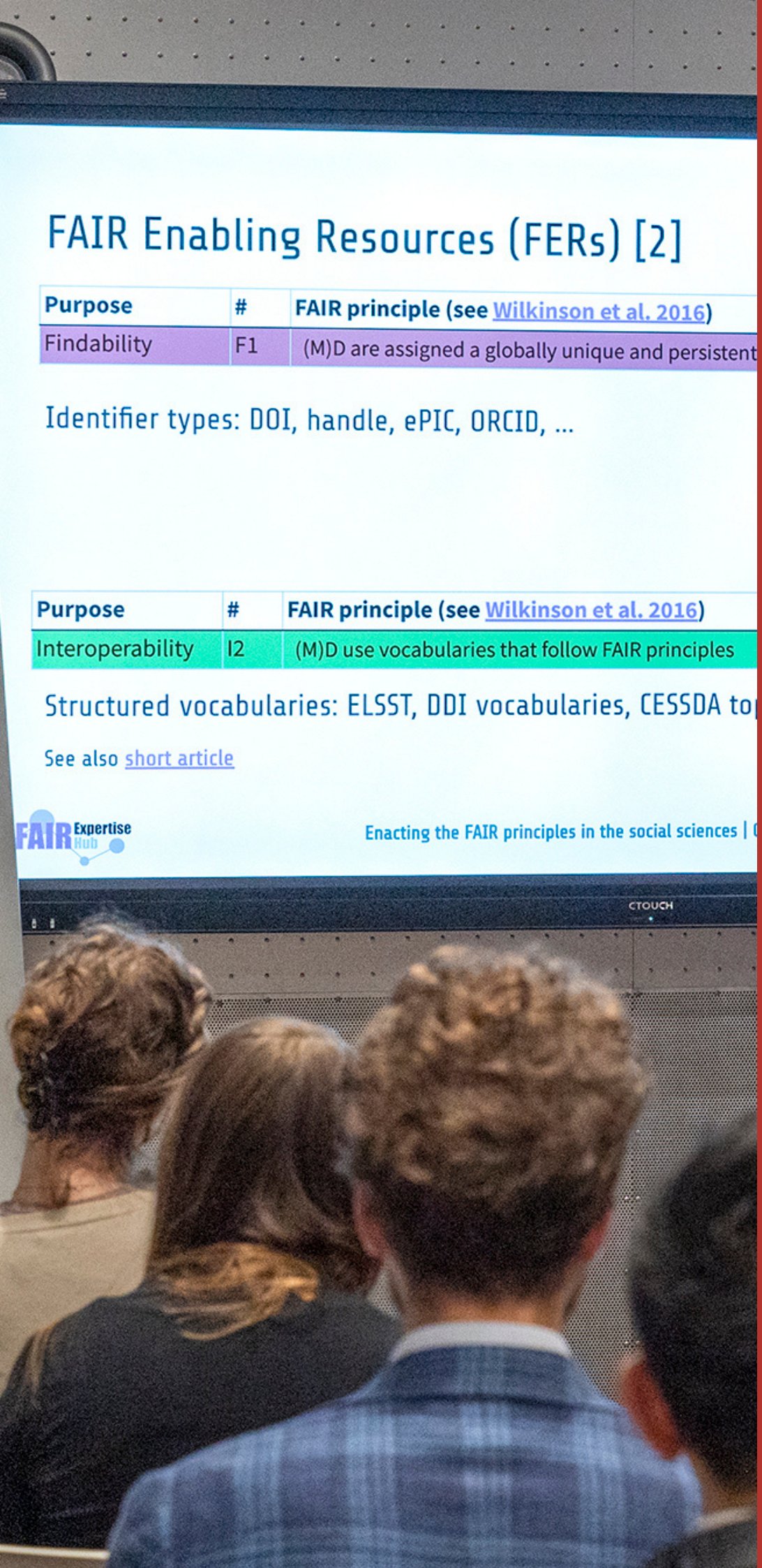
In the life sciences, the personal data in question is clinical data provided by hospitals, other health care practitioners, and public health data providers. In the social sciences there is a greater reliance on the use of TREs is in combination with administrative data and generally provided via a National Statistical Office.

This means that the social science infrastructure for TREs is almost exclusively developed at the national level, and there is currently very limited standardization in the implementation of these TREs across National Statistical Offices. The largest initiative in this area comes from the International Data Access Network (IDAN). However, this is sustained directly by just five data providers and does not have a framework for interoperable TREs.

We recommend that IDAN be radically scaled up and to provide TREs that are recognized as trustworthy by a large number of member states. This could include providing infrastructure for commercial and non-governmental organizations willing to provide secure research access to data they possess. This would not only increase the access to data for researchers but would also enable cross-national research with such data. Eurostat already provides a framework for the harmonization of administrative data, but this could be expanded and better supported by a network of TREs with a specific focus on supporting research.

Investments in IDAN would have two crucial benefits for the European Research Area. Firstly, they would enrich the existing ERICs and complement survey data collection investments. It is currently possible to link SHARE data to administrative data in eight countries, but it is currently impossible to analyse this data across the eight countries simultaneously. This does not have to be the case.

Secondly, TREs are pervasive in the North and West of Europe with the UK, Sweden, Norway, Finland, Denmark, and the Netherlands being the leaders in this field. However, this infrastructure is flexible and transferable. National Statistical Offices in the South and East of Europe could utilize this network without the high initial investments in infrastructure. The creation of a robust and flexible TRE network would therefore represent a radical transfer of resources from the North and West to the South and East and enrich the whole of the European Research Area.



FAIR

PAGES
8-9

RECOMMENDATION #4: CREATE PROFESSORIAL CHAIRS TO SUPPORT FAIR

Since the FAIR data principles were set out in 2016 (Wilkinson et al., 2016), there has been a large amount of activity and investment in ensuring that data across the sciences is FAIR and machine-actionable. The social sciences are no different.

The European Social Sciences are exceptionally lucky as CESSDA, the European consortium of national data archives, provides a natural home for the implementation of FAIR within the field. This has not been the case in many other fields, and this is reflected in CESSDA's flagship role in many European FAIR Initiatives. CESSDA leads the SSHOC-EU community and helps to address FAIR implementation issues specific to social science data providers and as the necessary expertise across its network to drive this change.

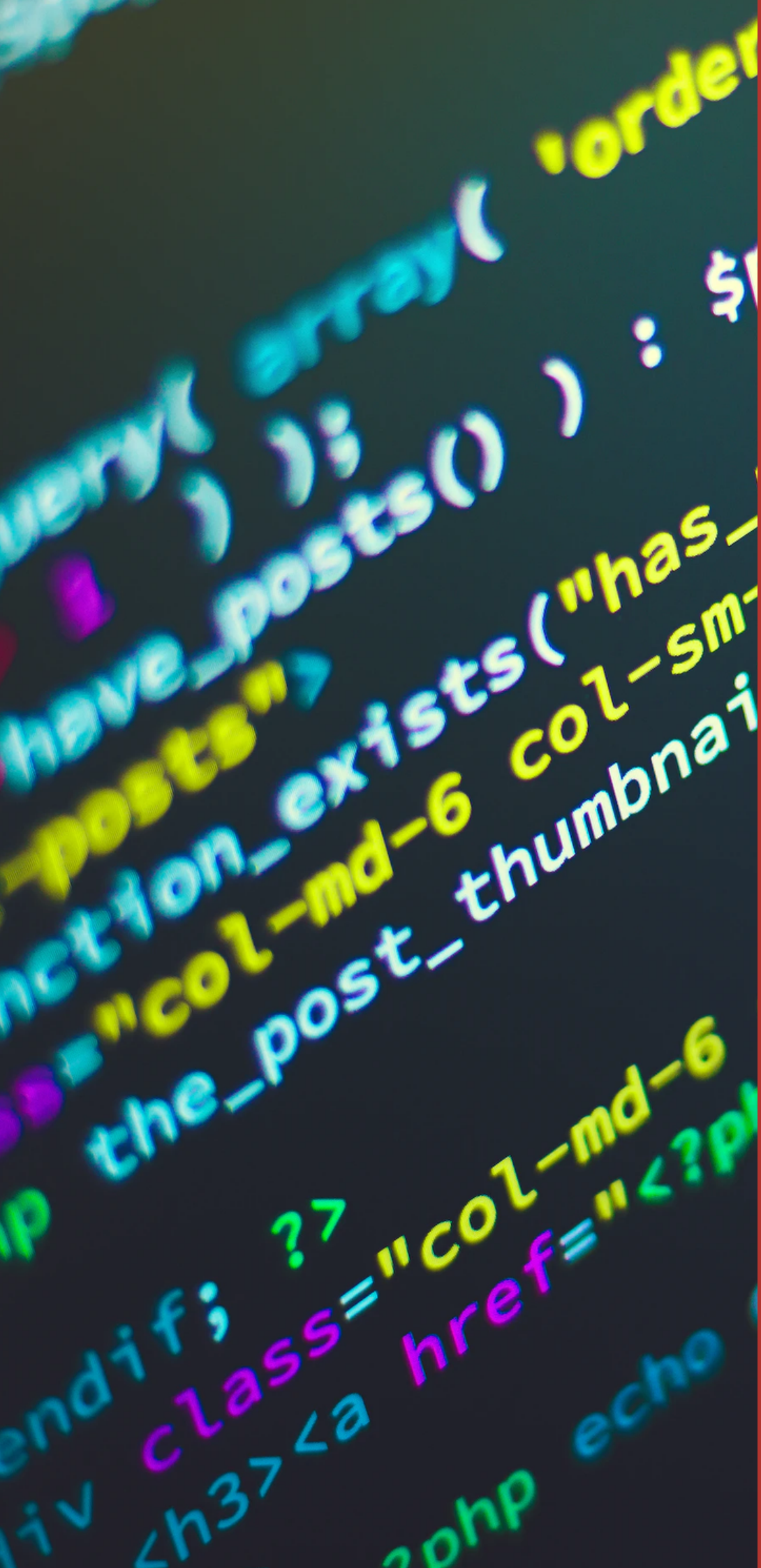
However, as with many other fields of research the key challenge in the social sciences is engaging the research community in the FAIRification of existing vocabularies and ontologies so that they can be used to enrich and link data from across the domain. As a matter of fact, many of these vocabularies and ontologies are widespread in the field in the form of codebooks and classifications. Yet they lack machine-readable representations and proper identifiers for terms. Therefore a disconnect remains between FAIR infrastructure and the community of data providers and scientists that are actively conducting research. To fill this gap, leaders in the field need to be supported in coordinating and enacting the FAIRifciation of their field.

CESSDA has set itself the goal of developing its strategic partnerships with the Social Science Research Community by 2027 but the commitment from the Research Community itself is more ambiguous and within the research ecosystem itself, FAIR remains an unfunded mandate for scientists collecting or analyzing data. The recruitment of data stewards in this area only goes some way to solving this issue as these positions are often classified as support rather than scientific positions.

To correct the imbalance and asymmetry in the commitments to FAIR implementation we recommend that scientific associations and universities support and encourage the creation of Chairs that are dedicated to leading FAIR implementation for specific domains.

These FAIR Leaders would be responsible for ensuring that their field not only generates new knowledge but also does so in a way that is FAIR and machine-actionable. By establishing only a few laureates across Europe, the profile for FAIR implementation work in the scientific community would be greatly elevated and would reaffirm that a commitment to FAIR is a commitment to science.

These Chairs would allow their host institutions to take a leading role in the strategic partnerships that CESSDA seeks and help make FAIR a reality for the domain.



SOFTAWARE

RECOMMENDATION #5: CREATE A NETWORK OF EXCELLENCE FOR SOCIAL SCIENCE SOFTWARE

The explosion and diversification in the data available to social scientists only represent part of the challenge faced by researchers today. As well as ever-changing and evolving data, the tools used to analyze that data and the models that are generated from that analysis are also ever-changing, and currently, the Social Science community is relatively underdeveloped in coordinating and managing these rapid changes.

The maintenance of crucial R, Python, or Julia packages, upon which a whole sub-field can rely, is often entirely left to the goodwill of the original developer. There is little support and coordination for this work and in time this will erode the replicability and standardization within the field at a time when continuity is crucial. Even in instances when a tool or model is heavily supported from an institutional perspective, the broader alignment and integration across the community is not actively promoted and it is left to the resources of individual institutions to support vital elements of social science infrastructure. A crucial example of this is EUROMOD.

EUROMOD is a model of tax and benefit systems of the European Member States that can be used to produce policy simulations and help policy-makers better understand the impacts of fiscal policy. It is, however, sustained by the Joint Research Center of the European Commission. It is a powerful tool for the research community, but the development of features that would be of benefit to the wider research community, such as a Python or R package, remain out of the purview of EUROMOD given that these developments are not a priority for the JRC.

There are many similar examples, and most codebases in the social sciences have access to far fewer resources than EUROMOD. For example, the Comparative Panel File is an open-source codebase that harmonizes the Longitudinal Panel Studies from across the globe. It is a wonderful idea and resource, but it will need to be maintained, developed, extended, and integrated within wider initiatives as they develop. Part of the challenge here is that these tools and models are often developed by scientists who are not equipped with the training and capacities to build sustainable software that is relatively future-proof. Many of the first developers of such tools are also now heading towards retirement and there is a sustainability crisis developing for many of these tools.

To address these challenges we propose a network of excellence for the development, deployment, and maintenance of research software in the social sciences. There is already an International Research Software Engineers community that coordinates efforts across domains. We argue that a specific network is needed, however, that is explicitly focused on the challenges faced by the social sciences. Specifically, this network should 1) attempt to coordinate efforts in the development, maintenance, and archiving of software and tools in the social sciences, 2) Train PIs of projects that generate software and tools in Software Development Skill sets, 3) Act as a host of last resort as the first great generation of software developers in the social sciences heads towards retirement. Europe has a large amount of pre-existing expertise in this area at places such as GESIS, the UKDA, FORS, and SoBigData Europe. These efforts should be brought together in this network of excellence.

EXPERTISE

PAGES
12-13

RECOMMENDATION #6: ESTABLISH AN EU-WIDE CSS GRADUATE SCHOOL

The speed of developments with regard to data, methods, and tools available to social scientists can be both intimidating and exhilarating for young scholars starting to conduct their first research projects. However, the biggest constraint on the uptake and utilization of these opportunities is training and access to the right knowledge resources.

Young scholars today not only need to familiarize themselves with sociological, economic, or sociological theory they also have to absorb a fast-changing range of advanced methods such as machine learning, neural networks, and natural language processing. This has always been one of the exciting challenges of a career in science but these new methods increasingly require scholars to develop skill sets in areas that draw heavily from other departments, such as software development, complex data management, data governance, or computational resource management.

These skill sets are hard to access and develop amongst young social scientists, and they are largely left to their own devices when it comes to translating and applying these. Even the most advanced computational social science faculties in Europe struggle to provide sufficient training and expertise in the eclectic and diverse range of methods that are required in modern social research.

Europe does, however, have a rich Erasmian tradition that can help address this by supporting the mobility of young social science scholars and providing the opportunities to access the expertise and training required. The European Consortium of Sociological Research and the European Consortium for Political Research already offer graduate courses on advanced topics, and the European Doctoral School of Demography provides first year PhD Students with the eclectic technical skills needed to conduct cutting-edge demographic research. There are also a range of excellent methods for summer schools and trainings such as the Essex Summer School, UPF, Ljubljana, and the Summer Institute for Computational Social Science.

There are however significant gaps in training, specifically in skills that will help develop the field more broadly. These include basic skills in FAIR data management, project management and development, science communication, advocacy, and legal and ethical dimensions of research. Infrastructures such as CESSDA and SoBigData have made efforts in this area but broader support and coordination is needed.

We propose to strengthen and extend these current initiatives through the creation of a European Computational Social Science Graduate School. This Graduate School would aim to provide fully accessible courses that provide ECTS accredited training by leaders in their respective fields. These courses could be integrated with and extended from existing offerings but with a view to provide greater interdisciplinarity, clearer accreditation, and greater capacity. The European Doctoral School for Demography provides a model for how this could be developed and implemented.

GOVERNANCE

PAGES
14-15

RECOMMENDATION #7: BUILD A FEDERATED EUROPEAN INFRASTRUCTURE CONSORTIUM FOR SOCIAL SCIENCE

The fault line that runs through the previous challenges is the relationship between research and research infrastructure. With all research infrastructures, it is vital that they lead the research community by implementing new technologies, enabling new methods, and providing access to new data. Ultimately however, research infrastructures must be responsive and accountable to the field which they serve.

Currently within the European Framework, this connection is often lost. FAIR investments are made only in infrastructure and so the research initiatives are lacking. Large-scale data collections are created and then isolated from the research community they were designed to serve. New tools and methods are enabled, but a lack of coordination with Universities and training centers mean that the research community is ill-equipped to exploit them. These investments are of exceptional valuable but far more could be gained from them with greater coordination and alignment across the social sciences.

This issue is in large part due to the design of ERICs themselves, as funding agencies and ministries constitute the board of these organizations and are rarely best placed themselves to judge whether an infrastructure is meeting the needs of scientists. This is generally why funding agencies send funding applications out for peer review rather than evaluating them themselves. In ERICs however, funding agencies are ultimately responsible for evaluating performance. This represents a considerable asymmetry in scientific knowledge between the team running the ERIC and those evaluating the performance and the scientific aims of the ERIC are therefore largely set by the executing team.

We propose to fix this misalignment through the creation of a Federation of European Social Science Infrastructures. This Federation will be constituted by two sets of organizations. On one side it will include infrastructures or service providers and specifically the Social Science ERICs and ESFRI members: ESS, SHARE, CESSDA, GGP, GUIDE, SoBigData Europe and any future social science infrastructures within the ESFRI framework. On the other side, there would be field organizations and specifically any European wide organization with more than 2,000 paying members from at least 20 countries. Examples of such organizations include but are not limited to the European Consortium for Political Research, European Consortium for Sociological Research, the European Association for Population Research, the European Economic Association, the European Federation of Psychologists Association, the European Survey Research Association and the European Communication Research and Education Association.

These field organizations would pay a small membership fee to be part of the Federation and in return would be able to nominate a member of a small Federal Council. This Federal Council would then be responsible for the evaluation and coordination of European Social Science Infrastructure in Europe, ensuring that Research Infrastructure serves the needs of scientists and leads to ground breaking research. This would alleviate the burden on national funding agencies of evaluating ERICs and would make the investments in European Social Science Infrastructure accountable to the field.

PHOTO CREDITS

© NicoElNino - stock.adobe.com, 2022

© Pietro Jeng - Unsplash, 2017

© Shahadat Rahman - Unsplash, 2019