The Greek Review of Social Research

Vol 163 (2024)

163 Special Issue: The computational turn in social sciences. Editors: Yannis Skarpelos, Kostas Karpouzis, Martha Michailidou

Challenges and opportunities for the re-use of New Data Types (NDTs) in a changing landscape

Dimitra Kondyli, Nicolas Klironomos

doi: 10.12681/grsr.38517

To cite this article:

Kondyli, D., & Klironomos , N. (2024). Challenges and opportunities for the re-use of New Data Types (NDTs) in a changing landscape. *The Greek Review of Social Research*, *163*, 193–219. https://doi.org/10.12681/grsr.38517

The Greek Review of Social Research, 2024, 163
Print ISSN: 0013-9696 Online ISSN: 2241-8512

Dimitra Kondyli[*], *Nicolas Klironomos*[**]

CHALLENGES AND OPPORTUNITIES FOR THE RE-USE OF NEW DATA TYPES (NDTS) IN A CHANGING LANDSCAPE

ABSTRACT

Over the past fifteen years, technology has contributed to the emergence of new types of data, particularly big data, influencing the methods of observation, study, and measurement of social phenomena from the perspective of the social sciences. The increasing digitization of social activities generates vast amounts of data that fuel contemplation about the way modern societies function. Additionally, factors such as the recent COVID-19 pandemic with mandatory social distancing have contributed to the creation of a favourable environment for the generation of new types of data, with an emphasis on big data. Within this ongoing transformation of the data landscape, we will attempt to pose questions related to the environment of Data Repositories/Research Infrastructures and the means/methods of addressing and managing these data. It appears that social research is shifting towards a more "data-driven approach," which requires new skills and capabilities at the intersection of the computational and social sciences. One of the major issues that arise is the potential for collaborations between data organizations and researchers/users of data to promote not only a culture of data sharing but also the reuse of such data. This work will be based on primary and secondary sources generated within the framework of research projects in collaboration with CESSDA ERIC (European Social Science Data Archives-European Research Infrastructures), as well as literature on the management of data from various sources, with an emphasis on their legal/ethical and technical aspects.

Keywords: *new data types, data repositories, data re-use, research infrastructures*

[*]Research Director at the Institute of Social Research of the National Centre for Social Research, President of the SoDaNet Steering Committee, National Representative of SoDaNet at CESSDA - ERIC, e-mail: dkondyli@ekke.gr

[**]Political Scientist, Scientific Associate of the National Centre for Social Research, e-mail: nklironomos@ekke.gr

Received: April 26, 2023 Accepted: January 17, 2024 Published: July 30, 2024

Δήμητρα Κονδύλη[*]*, Νίκος Κληρονόμος*[**]

ΠΡΟΚΛΗΣΕΙΣ ΚΑΙ ΕΥΚΑΙΡΙΕΣ ΓΙΑ ΤΗΝ ΕΠΑΝΑΧΡΗΣΗ ΝΕΩΝ ΜΟΡΦΩΝ ΔΕΔΟΜΕΝΩΝ ΣΕ ΕΝΑ ΜΕΤΑΒΑΛΛΟΜΕΝΟ ΤΟΠΙΟ

ΠΕΡΙΛΗΨΗ

Τα τελευταία δεκαπέντε χρόνια η τεχνολογία συνέβαλε στην εμφάνιση νέων τύπων δεδομένων, ιδίως μεγάλων δεδομένων, που επηρεάζουν τις μεθόδους παρατήρησης/μελέτης/μέτρησης των κοινωνικών φαινομένων από την πλευρά των κοινωνικών επιστημών. Η αυξανόμενη ψηφιοποίηση των κοινωνικών δραστηριοτήτων παράγει τεράστιες ποσότητες δεδομένων που τροφοδοτούν τον προβληματισμό για τον τρόπο λειτουργίας των σύγχρονων κοινωνιών. Επιπρόσθετα, παράγοντες όπως η πρόσφατη πανδημία Covid-19 με την αναγκαστική κοινωνική αποστασιοποίηση συνέβαλαν στη δημιουργία μιας επιπλέον ευνοϊκής συγκυρίας έτσι ώστε να εισέλθουν πιο εντατικά στο προσκήνιο νέοι τύποι δεδομένων με έμφαση στα μεγάλα δεδομένα. Στο πλαίσιο του διαρκούς μετασχηματισμού του τοπίου των δεδομένων, θα επιχειρήσουμε να θέσουμε ερωτήματα που σχετίζονται με το περιβάλλον των Αποθετηρίων Δεδομένων/ Ερευνητικών Υποδομών και τα μέσα /τρόπους αντιμετώπισης και διαχείρισης αυτών των δεδομένων. Φαίνεται ότι η κοινωνική έρευνα στρέφεται προς μια πιο "καθοδηγούμενη από τα δεδομένα προσέγγιση", η οποία προϋποθέτει νέες δεξιότητες και ικανότητες στο σταυροδρόμι των υπολογιστικών και κοινωνικών επιστημών. Ένα από τα μείζονα ζητήματα που αναδεικνύονται είναι η δυνατότητα συνεργασιών μεταξύ οργανισμών δεδομένων και ερευνητών/χρηστών των δεδομένων προκειμένου να προωθήσουν όχι μόνο μια κουλτούρα κοινής χρήσης δεδομένων, αλλά και επανάχρησης αυτών. Η εργασία αυτή βασίζεται σε πρωτογενείς και δευτερογενείς πηγές που έχουν παραχθεί στο πλαίσιο ερευνητικών έργων σε συνεργασία με το CESSDA ERIC (European Social Science Data Archives- European Research Infrastructures), καθώς και σε βιβλιογραφία αναφορικά με τη διαχείριση δεδομένων προερχόμενων από διαφορετικές πηγές με έμφαση στις νομικές/ ηθικές και τεχνικές πτυχές τους.

Λέξεις-κλειδιά: *νέοι τύποι δεδομένων, αποθετήρια δεδομένων, επανάχρηση δεδομένων, ερευνητικές υποδομές*

[*]Διευθύντρια Ερευνών στο Ινστιτούτο Κοινωνικών Ερευνών του Εθνικού Κέντρου Κοινωνικών Ερευνών, Πρόεδρος της Διοικούσας Επιτροπής του SoDaNet, Εθνική εκπρόσωπος του SoDaNet στη CESSDA - ERIC, e-mail: dkondyli@ekke.gr

[**]Πολιτικός Επιστήμονας, Επιστημονικός συνεργάτης του Εθνικού Κέντρου Κοινωνικών Ερευνών, e-mail: nklironomos@ekke.gr

INTRODUCTION

This paper will attempt to underline the importance of New Data Types (NDTs) for empirical social research, further exploitation and re-use. An essential component of the current work aims at connecting research outcomes raised by the use of new data from the Research Infrastructures perspective. In the last fifteen years, empirical social research has been impacted and facilitated by specialised institutions (the so-called Research Infrastructures/Data Repositories/Data Archives), which rendered social empirical data available for re-use to the wider community. By providing datasets ready for re-use and reproducibility, Data Repositories act at the same time as "guardians" of the data life cycle, per se as well as of the collective memory and specific social instances.

The rapid development of technology and software tools along with the creation of a unique European Research Area (ERA)[1] favoured the development of Research Infrastructures, a driving force for the promotion of scientific and responsible research and data services for the benefit of European societies (ESFRI 2006; 2008; Chou, 2014; Ulnicane, 2015). Certified Data Repositories[2] create stable digital environments that promote and facilitate the re-use of data following documentation procedures. Along with the development of certified Data Repositories and Data Archives came the development of Open Science principles that encouraged the production and accumulation of all data types and formats, including NDTs. What is Open Science and how does it contribute to the aforementioned objectives? "*Open Science is the new standard of practices, means and collaboration for producing and distributing scientific output and research results, with a direct scientific, economic, and societal impact"* (Athanasiou et al., 2020). It aims to build an ecosystem in which science is more cumulative, data-

1. For more information on The European Research Area (ERA): https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/european-research-area_en

2. Certified and trusted Repositories or Data Archives have been awarded a reviewing process based on specific requirements in order to comply management operations regarding data management, access policies, compliance with FAIR principles etc. with international standards. Organisations like CoreTrustSeal (CTS) https://www.coretrustseal.org/apply/ or World Data System (WDS) provide reviewing process that conclude to certification awards. Within the frame of Horizon 2020 as well as the current Horizon Europe the European Commission underlines the importance of certified repositories which support open access where possible regarding data, related metadata and code to be deposited. More at https://www.openaire.eu/find-trustworthy-data-repository-certified-repositories

driven and universally accepted. From vision to implementation, Open Science principles promote more responsible research by providing easier access to produced knowledge, reliable data, training tools and education courses for researchers to build adequate data skills, and critical thinking. This process of knowledge democratisation is open to every citizen. In the ERA and related research funding mechanisms, Open Science constitutes a critical priority, as it encourages open access to research outcomes and developments.

Data sharing and open access not only facilitate equity in data access, narrowing the divide between the "*data rich*" and the "*data poor*" (Boyd & Crawford, 2012; Metzler, et al., 2016) but can also be conceived as a way of promoting the reproducibility of research. Reusability allows researchers to build upon and continue the re-analysis of research that has been produced by primary data authors through the process of verification of results and reproduction of new outcomes (Thanos, 2017). An aspect of research reproducibility from the repository's perspective is the handling of replication materials such as data, metadata and code (Kondyli et al., 2024). Because of Open Science, scientific Journals and publishers require the replications of data analysis which are included in positively reviewed and accepted articles. Those replications are preferably hosted in recommended Data Repositories. The stake for Data Repositories is great because they can support researchers in handling replication analysis, cooperating with Journals to design their data policies and finally providing any potential user with the replication analysis for re-use[3] (Sawchuk & Khair, 2021).

DEALING WITH NDTS FROM THE REPOSITORIES' PERSPECTIVE

In recent years, there has been a significant increase in NDTs in the study of social phenomena. This type of data is often secondary data, which may be used by social researchers for specific research other than the purposes of the primary collection, or any operational or even commercial purposes (Boté & Termens, 2019; Hox & Boeije, 2005). These NDTs include social media data, all kinds of digital information about the human condition, data derived from digital sensors, financial transactions as well as administrative

3. More information about the implication of Data Repositories in replication services in the CESSDA Data Archiving Guide (DAG) available at https://dag.cessda.eu/About. The chapter on replications services is forthcoming.

records (OECD, 2016). In the literature, these types of data have also been characterized as "*found data*" because they were not initially collected for research (Harford, 2014; Hemphill et al., 2022).

Table 1 provides a typology of NDTs published by the OECD (2013), which has been further used for a recent research project[4]. Thus, working questions and arguments developed in this paper are inspired by the following categorisation of NDTs. However, due to the limitations of the present work, we would like to stress that a particular emphasis will be given to specific categories of NDTs, to the conceptual connection with computational science as well as good practices of CESSDA and worldwide repositories in the management of NDTs. Thus, taking into consideration these elements, emphasis will be given to some commonly managed NDTs out of the following categories, including data derived from internet use (i.e., social media data), satellite and aerial imagery data (remote sensing data), as well as health data that experienced a considerable increase due to the recent Covid-19 pandemic.

The increase of NDTs provides material for social empirical research on the one hand, whilst on the other raising several issues related to the research data culture from the point of view of both researchers and Data repositories. Social science researchers familiar with data are also users of data repositories services mainly in countries where data sharing culture is highly valuable (Stuart et. al., 2018; Digital Science, 2017). Data Repositories have contributed substantially to the development and promotion of a data-sharing culture. In addition, several factors related to the importance of research citations, the re-use of reliable datasets, the evolution of technological research tools the high cost of comparative quantitative surveys and continuous efforts for research funding, enhanced the interplay between two of the components of the research ecosystem worldwide. Thus, Data Repositories have provided reliable services to various research communities in order to produce responsible research outcomes.

Technological developments also have an impact on the variety of data available through the web, which is linked to the development of the data revolution or data evolution. This brings us to an "emerging phenomenon"

4. CESSDA ERIC Agenda 21-22, Tasks 21-22 Widening & Outreach Pillar: Task 2 Survey on Researchers Needs and Widening the Perimeter of Data. Deliverable Overview and summary of existing outputs (inside and outside of CESSDA) on NDTs, elaborated by B. Kleiner, D. Kondyli, N. Klironomos, L. Bishop, M. Vavra, Th. Cizek and Y. Leontiyeva.

Table 1: *NDTs typology by OECD (2013)*

Category A: government transactions	Individual tax records, Corporate tax records, Property tax records, Social security payments, Import/export records
Category B: government and other registration records	Housing and land use registers, Educational registers, Criminal justice registers, Social security registers, Electoral registers, Population registers, Health system registers, Vehicle/driver registers, Membership registers
Category C: commercial transactions	Store cards, Customer accounts, Other customer records
Category D: internet usage	Search terms, Website interactions, Downloads, Social media data, Blogs; news sites
Category E: tracking data	CCTV images, Traffic sensors, Mobile phone locations, GPS data
Category F: satellite and aerial imagery	Visible light spectrum; Night-time visible radiation, Infrared; radar mapping
Category G: health data	MRIs, ultrasounds, neuroimaging data, patients' records", CT scans, X-rays
Category H: other data types	All new data types other than those mentioned above.

within the world of data and related actors. Researchers familiar with these NDTs are often able to use IT tools and programming languages that are appropriate for conducting their research for their specific purposes. By doing so, they can navigate and conquer as many megabytes of data as they want to easily, efficiently and rapidly do research and make relevant publications, so a turning point in the established data culture may be produced. On the other hand, Data Repositories face the dilemma of attracting these young generations of highly skilled social science researchers while needing to overcome barriers raised by the management of NDTs.

SOCIAL SCIENCE RESEARCH COMMUNITIES AND DATA REPOSITORIES: A COMPOSITE INTERPLAY

In the "*World Social Science Report 2016*", Mike Savage (2016) argues that NDTs constituted mainly by big data are now an integral part of new types of data collection in today's social sciences research. Thus, whether we like the idea or not, social science research should create a favourable environment and fertile ground for exchange in terms of methodology and use. In empirical social research, research data have mainly been produced by either quantitative or qualitative surveys. Social scientists should now reflect and further elaborate on methodological acquis of a scientific status quo (surveys, questionnaires, specific sample populations, etc.), which capture social representations determined in time and conjecture to reproduce related analyses beyond the legitimating jurisdiction of their disciplines. These data analyses provided by the huge accumulation of data can motivate social scientists to study social phenomena using data that were not originally collected for research, such as social networking data, data derived from online information mining, data derived from digital sensors as well as administrative records (OECD, 2013). It could then be argued that social sciences cannot remain outside the emerging body of knowledge within this data landscape and its possibilities.

For Lazer & Radford (2017), the most compelling sociological research in the twenty-first century will not be big data but a fusion of data sources related to important questions. The existing plurality of data types and forms provide new insights into human behaviour and aspects of contemporary life, which makes them particularly attractive to social science researchers. The monitoring of large samples produced by huge amounts of high-detail data over long or short periods is relatively easy and often acquired at a lower cost than survey data (Ackland, 2013; Conrad et al., 2019; Karpf, 2012; Kosinski et al., 2015). Conducting social science research based on NDTs also requires specific skills for social science researchers that presuppose basic or solid knowledge of data science or computational skills, beyond conventional social sciences curricula (Giglietto & Rossi, 2012).

Along with computational skills, researchers must also be aware of the legal and ethical possibilities and limitations of using NDTs. The complex and rich data landscape has many facets. The antagonistic research environment and the intensive efforts for publication within the academic community operate as a catalyst, constantly pushing researchers to find new sources of research material (the so-called "*publish or perish*"). Among these sources, data-driven research remains a key element of

this production. Furthermore, the variety and diversity of such data allows researchers to simultaneously capture instances of social reality to restructure and reuse them in the way they believe best meets their working hypothesis. Usually, a paradox occurs that restrains the re-use of such research outcomes by interested third-party users. Focusing on the three categories of new types of data already mentioned, it becomes clear that in the vast majority of cases, these data are mainly collected or produced by other data holders or made available via digital platforms. Contrary to survey data, researchers do not fully comprehend how the data have been produced and, more importantly, how and whether these data, which relate to information about people's views, attitudes and characteristics, comply with ethical and legal requirements.

The General Data Protection Regulation (GDPR) came into force in time to respond to the abundant and changing circulation of information and the complex data landscape. The GDPR and the current ethical framework concerning the circulation of information and data protection have set limits, creating at the same time opportunities to normalise/regularise the data landscape for research purposes. To an extent, Research Infrastructures and/or Data Repositories deal with the re-use of NDTs via their mechanisms and procedures. Based on a recent report (Kleiner et al., 2022), several Service Providers of CESSDA collect, store and disseminate for re-use NDTs. The work attempts to depict via an extensive review of the literature and a web survey the lessons learnt by the archival practices of CESSDA Service Providers (SPs) regarding NDTs, as well as to reflect upon modes of cooperation and transfer of know-how among them to meet researchers' needs and deal with relevant challenges. Open Science principles have mobilised the data landscape, increasing the scope of stakeholders involved in the research process. Funding frameworks *i.e.*, Horizon or Journals, promote data-sharing practices and policies from a research transparency perspective, aiming at responsible research per se. Funding frameworks of national or EU provenance, aim for the widest dissemination and publication of produced data of research projects, while Journals adopt data-sharing policies to make the publication process more transparent by reviewing and controlling the data of a publication. Research Infrastructures and trusted Repositories can be or become the designated places to achieve these objectives and aims.

Lately, these data institutions have been faced with challenges and opportunities that affect their mission and daily routine. Thus, they can be appropriate actors and key players in the composite interplay by

mobilising qualified staff and evolving processes and practices. Beyond the management of quantitative survey data and qualitative research data, RIs and Data Repositories are called upon to deal with more complex forms of data and data formats. Additionally, and beyond technical challenges, Data Repositories have to efficiently solve legal and ethical barriers to serve their respective research communities in the long run. Preconditions for processing and managing NDTs are the highly skilled personnel of the Data Repositories as well as adjusted processes and practices to capitalise on previous experiences and build capacity for the new research and data environment (Kondyli & Linardis, 2019). In other words, these preconditions are the essence of Data Repositories' work and perspectives.

ADJUSTING PROCESSES AND PRACTICES FOR NDTS RE-USE?

As different types of NDTs have different characteristics, they create different types and intensities of challenges for research infrastructures and data repositories. As part of subtask 2 of CESSDA's research project, "Widening & Outreach Pillar: Task 2 Survey on Researchers' Needs and Widening the Perimeter of Data", a survey was conducted among CESSDA's Service Providers[5] (SPs) about hosting NDTs in the individual repositories. The purpose of the survey was to ascertain the needs and challenges for archives related to NDTs as well as how SPs can support each other and become better equipped to cope with the changes that hosting NDTs brings to their archives. The survey was designed by the subtask team from the CESSDA SPs: ČSDA,[6] SoDaNet,[7] GESIS,[8] and FORS.[9] The operational definition of NDTs used for the survey and provided to respondents was the following: *"any kind of data that challenges and presents particular difficulties for our traditional archiving practice"*. In asking about the

5. The SPs that participated are 24: 21 members/observers of the CESSDA consortium and 3 partners.

6. Czech Social Science Data Archive (ČSDA) is a national resource centre for social science research, which acquires, processes and archives datasets from Czech and international social research. For access to the data hosted in ČSDA: https://archiv.soc.cas.cz/en/

7. Social Data Network (SoDaNet) is the research infrastructure of Greece for the social sciences. For access to the data hosted in SoDaNet: https://datacatalogue.sodanet.gr

8. GESIS Leibniz Institute for the Social Sciences, based in Germany, is the largest European infrastructure institute for the social sciences. For access to the data hosted in GESIS: https://www.gesis.org/en/home

9. Swiss Centre of Expertise in the Social Sciences, is the national centre of expertise in the social sciences. For access to the data hosted in FORS: https://forscenter.ch

different NDTs archived with SPs and their corresponding challenges, the typology that was employed is the same as we presented above in Table 1 from OECD (2013). Considering the categories of NDTs that we examine in this paper (internet usage data, satellite and aerial imagery data, tracking data and health data), some interesting conclusions emerge.

Hosting NDTs is not so extensive in the CESSDA SPs

As both the use and availability of these data have become practically possible relatively recently, most of the data available from CESSDA SPs do not fall into the categories of NDTs. On the other hand, this is changing rapidly and in particular, the "larger" and more "established" Archives have at least one or all of the NDTs we mention, while smaller Archives follow, as we will see from the SoDaNet example in a subsequent chapter.

NDTs that researchers request to archive

As data archives and research infrastructures are not themselves data producers but act as bridges between data producers and secondary users and research communities, the data hosted are those requested by researchers. CESSDA SPs were asked about the requests for hosting and

Figure 1: *Wordcloud of "Specific NDTs that researchers have been requesting to archive". Answers from CESSDA SPs that participated in CESSDA's "Survey on Researchers Needs and Widening the Perimeter of Data". Wordcloud produced by the authors*

documentation they have received regarding NDTs. As we can see from Figure 1, these requests are for the most "popular" types of NDTs that we are studying here, such as Social Media Data, Internet Usage Data, Health Data and Tracking Data as well as more specialized data types that can be included in the above categories such as MRI Data (Health Data), Internet Behaviour Data (Tracking and Internet Usage Data) etc.

LEGAL-ETHICAL AND TECHNICAL ISSUES OF NDTS FOR DATA REPOSITORIES

NDTs offer researchers great opportunities for interesting and challenging research. The regulatory framework for the use of data, particularly when they belong to the category of so-called ''personal'' data in the context of NDTs and data repositories, is defined and must comply with the GDPR. Research usually involves the processing of personal data. According to the European Data Protection Supervisor, "*A good privacy notice should tell you who is collecting your information, what it is going to be used for, whether it will be shared with other organisations*". The above applies to all types of data used in empirical social research. Given the complexity of the sources from which the data is obtained, particularly concerning personal data protection issues such as informed consent,[10] there is often an inability to effectively track the entire data collection process. However, given the complexity of the sources from which the data is obtained, *i.e.*, the inability to keep track of the whole data collection process, particularly concerning issues of identification of personal information. In addition to the identification of individuals, another restriction is that the individual data subject providing the information must have given his or her prior consent for further sharing and publication of the content of his or her utterances.

A significant part of the literature focuses on the responsibility of researchers to take all necessary measures and steps to protect the identification of data subjects, together with the informed consent of data subjects regarding the further use of their information. Ethics of research cover the whole spectrum, from collection to storage and usage of large-scale data, the biases inherent in the dataset itself, consent of data owners,

10. Processing of personal data in reliance on a legal ground specified in Article 6 GDPR informed consent (a) refers to data subject rights to be informed (informed consent), the right to be forgotten (erase information) or the right to withdraw consent (change of opinion)

potential risks and benefits arising from the analysis of the data and, last but not least, transparency. Salah et al. (2022) give the example of biometrics and tracking applications in the research area of migration and mobility. Tracking applications of people in movement allows governments and other stakeholders to identify and track people across large geographical areas without providing any possibility for people to know for how long for what and how governments are tracking their trajectories or if and how various agencies share these data or to react. The tools and methods of computational sciences fill in this kind of research, the outcomes of which can be transformed into policies.

The data centres operating in many European data archives or statistical offices ensure that both requirements are implemented by hosting the data in a secure environment and training researchers to process this type of big data. Beyond secure data centres, data repositories have a lot to teach and assist researchers, when it comes to dealing with ethical and legal constraints (Bishop, 2017). Until now, the know-how of Data Repositories in dealing with complex data and NDTs came mainly from the management of administrative data and in recent years from social media data that experienced massive growth in the previous decade (Mannheimer & Hull, 2018). The relationship between public and private is exhibited in this type of data, making its possession more complex and creating legal constraints at different levels. As an example, censuses and microdata of Statistics Bureaus or Official Statistics can be identifiable data when linked with administrative records, or data derived from economic activities at the lowest regional level etc. (Desrosières, 2005). Thus, sometimes this kind of data constitutes confidential data that cannot be considered as public-use data outcomes, which are usually processed, aggregated, and anonymised. They have been elaborated from initially confidential data, but the level of processing does not allow the identification of data subjects (Lagoze et al., 2013). The authors argue that trusted Data Repositories can undertake safely the processes of ingestion, curation and necessary modifications of both data and metadata to render them suitable for re-use.

Another possible impediment to the reuse of NDTs may be the origin and purposes of the data generated, i.e., the distinction between data funded by the private sector for business purposes and data funded by the public sector for the public interest. There may be relationships between them, which sometimes create complex chains so that it can be very difficult to trace the origins of a particular dataset. Where the particular dataset is not accompanied by unique standardised schemas i.e., persistent

identifiers (PIDs)[11] or complete metadata, the identification of preexisting relationships and origins of the data can be difficult to understand. Data Repositories can offer solutions/answers to many of these issues.

CESSDA Repositories such as UKDA[12] or GESIS curate the content of social media data by holding only tweet and user IDs for them to be retrievable, with no assurances that identical data will be produced (see Good Practices chapter below). Given that tweet data might be deleted when another researcher would like to reproduce the analysis, an alternative to sharing Tweet IDs is to only share derived data.

Another component that poses technical issues and covers different aspects of the Data Repositories in terms of data management practices and capacity is the computational processing required for certain types of new data, such as social media data and sensor data. Both researchers and data professionals from different perspectives need to constantly upgrade their professional skills and expertise. Computational skills like programming, web scraping, and programming languages are necessary for research in social media and sensor data (Hemphill et al., 2022; Bastin & Tubaro, 2018). Data professionals had not integrated computational practices into the management practices concerning survey data until quite recently. The "computational turn in social sciences" offered challenges and opportunities to adjust and transform data management practices. Thus, the documentation and general management of these data by institutions require the enrichment of metadata standards, and knowledge of advanced techniques, i.e., web scraping, and managing codes or notebooks, depending on the design of the given survey. The discussion about skills relies upon the extended services of Data repositories and in particular well-known data institutions that the authors happened to be more familiar with, like CESSDA Repositories. The core mission is composed of four strategic pillars, namely Tools, Training, Trust and Widening & Outreach activities. The four pillars are centred around people and data that strengthen CESSDA Repositories to substantially contribute to a composite and challenging

11. Persistent Identifiers (PIDs) serve to uniquely identify a publication, dataset, or person. The metadata for these persistent identifiers can provide unambiguous links between persistent identifiers of the same type, e.g. journal articles citing other journal articles, or of different types, e.g. linking a researcher and the datasets they produced.

12. UK Data Archive (UKDA) based at the University of Essex is the lead partner of the UK Data Service, providing researchers with support, training and access to the UK's largest collection of social, economic and population data. For access to the data hosted in UKDA: https://beta.ukdataservice.ac.uk/datacatalogue/

landscape with management practices that attempt to respond, among other issues, to the management of NDTs.

Based on CESSDA's ''Widening & Outreach Pillar: Survey on Researchers' Needs and Widening the Perimeter of Data'', NDTs present some main challenges to research infrastructures and social data repositories.

1. Legal/ethical issues: Internet Usage Data, Tracking Data and Health Data seem to be the most challenging NDTs in legal, ethical, and/or data protection issues. As the institutional and legal framework for the use of data from internet sources has not been as lucid -especially in the past decade- as expected, this also affects the data repositories that disseminate and share these data for secondary use. Health data in its various forms (MRI/CT scans, patients' records etc.) brings similar challenges as it is by nature quite sensitive data, but an attempt was made to overcome this very quickly in the context of the pandemic crisis.

2. Technical issues: Khan et al. (2021) report that an important incentive for institutional repositories is supporting academic researchers who follow funder policies, but often lack the technical expertise available for discipline-specific large-scale repositories. For example, sometimes the NDTs that the researchers request to archive can be classified as Big Data due to their large volume, complexity of the existing archival system, etc. The size of datasets brings different technical solutions and challenges, as Uzwyshyn (2016) states. For medium to large projects, data may require dedicated back-end storage systems to create larger storage options (e.g., dedicated network space allocation, RAID etc.), while very large data sets and projects require collaboration with larger organisations to provide web services that are often too expensive for social data repositories (e.g., DuraCloud, Amazon Web Services etc.). Resource availability in general can affect repositories in many ways. Issues such as:

    a. Data cleaning,
    b. Availability of time, resources and know-how,
    c. Archive's access to the data,
    d. Adaptation to the existing metadata schemas.

    were mentioned to a small or greater extent by all institutions that participated in the survey and are relevant to all four types of NDTs

that we focus on in this paper and also related to the material, human and technical resources available to repositories and research infrastructures.

GOOD PRACTICES

Whereas the use of digital devices is co-constructing human agency and social interactions in unprecedented ways, bringing about a rethinking of the theoretical assumptions of social science methods (Rupert, Law & Savage, 2013), at the same time it poses challenges to research infrastructures and data repositories (King, 2011), especially in privacy, data protection rights (Politou et al, 2021) and ethics of sharing (Bauchner et al., 2016; Resnik & Elliott, 2016). In this chapter, we will present some good practices in documenting data files that can be categorized as NDTs, from certified and widely known social data repositories such as GESIS, UKDA, DANS[13], ICPSR[14] and SoDaNet. As we have already discussed, what we call NDTs include several types of data; we will focus on those associated with the footprint of daily internet users(Tracking Data and Internet Usage Data such as Social Media Data) as well as those that, due to their nature, were more difficult to process and require large computational resources to be utilised (Remote Sensing Data), but also another category of data that have grown a lot due to the recent pandemic crisis of COVID-19, namely health data.

SOCIAL MEDIA DATA

The daily use of social media by millions of users around the globe creates an immeasurable wealth of social interaction and, by extension, data primarily in unstructured textual format. This data can be made accessible to researchers through data mining techniques that provide the tools for researchers to analyse large, complex, and frequently varying social media data (Barbier & Liu, 2011). Hosting such datasets raises challenges for

13. Data Archiving and Networked Services (DANS) is the Dutch national centre of expertise and repository for research data. For access to the data hosted in DANS: https://easy.dans.knaw.nl/ui/browse

14. Inter-university Consortium for Political and Social Research (ICPSR) is a n international consortium of more than 750 academic institutions and research organizations based at the University of Michigan, USA. For access to the data hosted in ICPSR: https://www.icpsr.umich.edu/web/pages/ICPSR/index.html

Figure 2: *Documentation of the dataset for the project "TweetsCOV19 - A Semantically Annotated Corpus of Tweets About the COVID-19 Pandemic" in GESIS repository*

research infrastructures and data repositories, especially in the area of privacy and compliance with regulations governing personal data such as the GDPR. Research infrastructures and data repositories are increasingly applying anonymisation techniques to the data they host to secure their secondary use (Zhou, Pei & Luk, 2008). Concerning the process of anonymization, we will focus on the example of the documentation of the dataset for the project "TweetsCOV19 - A Semantically Annotated Corpus of Tweets About the COVID-19 Pandemic" hosted in the GESIS Repository[15] (Figure 2).

15. For access: https://doi.org/10.7802/2470

GESIS provides a standardized and robust metadata documentation, with the basic information available in the documentation of more ''traditional'' forms of research data. However, what we need to emphasise is the anonymisation process in the data file as well as what is provided to secondary users:

1. The IDs and usernames of the tweets are encrypted and the text itself is not provided.
2. "*Entities*" in which the tweets are categorised at the content level are provided in the data file.
3. Any "*Mentions*" (to other accounts) and "URLs" that may be present in the text are provided in the data file.
4. The Sentiment Analysis of the tweets in the form of scores is provided in the data file.

At the same time, clear and detailed instructions of how the data was obtained are given through documentation on a separate page referring to the specific research project. This enables secondary users to conduct their analyses on this massive dataset without violating the personal data of Twitter users under the GDPR.

TRACKING DATA

The plethora of digital devices (smartphones, wearables etc.) in daily use with mobile sensors opens up a wide horizon for the exploitation of behavioural observation data that previously would have been impossible or at least very expensive (Harari et al., 2017). Unsurprisingly, such data may be difficult to obtain and raise ethical and practical challenges for researchers and research infrastructures (Breuer et al., 2020). They can also be quite sensitive, as they essentially record many privacy-related aspects of human behaviour and condition: from data relating to a person's health (such as sporting activity, heart rate measurements, etc.) to the entirety of a person's movements and interactions. Thus, depending on the nature of the data, research infrastructures and data repositories may place restrictions on data availability (access levels, embargo periods, on-site availability etc).

In the examples we present from the ICPSR repository, we observe that in the case of ''*Monitoring High-Risk Sex Offenders with GPS Technology in California, 2006-2009*'',[16] the data is restricted. To be accessed, a

16. For access: https://doi.org/10.3886/ICPSR34221.v1

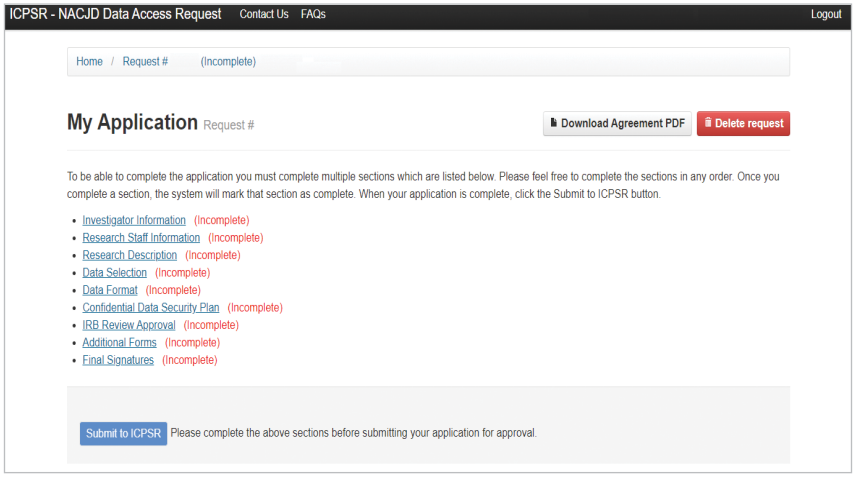


Figure 3: *Restricted Data Use Agreement in ICPSR*

Restricted Data Use Agreement must be completed, which requires the completion of detailed personal information which will then be submitted to IPCSR (Figure 3).

In the case of "*Mobile logdata of field trip learning in traditional village*,"[17] the provided data files are available but, as expected, they do not provide information about the individuals who participated, since they are anonymised. In both cases in the ICPSR repository, we do not observe any differentiation in the documentation of the research data, such as, for example, some specialised metadata fields related to the tracking data, seeing as the documentation is adequate.

It is increasingly common for digital trace data to be combined with more traditional data sources such as survey data. By doing so, the two data sources complement each other and give researchers better analytical capabilities (Silber et al., 2022; Stier et al., 2020). One such case is the Timing-and-Tracking Data of the "*Synaesthetic Engagement of Artificial Intelligence with Digital Arts and its Audience (AI TRACE)*"[18] research project, the data of which are hosted in the SoDaNet research infrastructure.

SoDaNet offers complete and standardized documentation at three levels/metadata modules (Linardis, Alexandris & Klironomos, 2022):

17. For access: https://doi.org/10.3886/E108741V1
18. For access: https://doi.org/10.17903/FK2/39NZQM

- Citation Metadata,
- Geospatial Metadata,
- Social Science and Humanities Metadata.

The data is anonymised so that the secondary user cannot link the questionnaire data and the timing-and-tracking data taken during the participants' visit to the museum with the actual persons.

REMOTE SENSING DATA

In recent years we have seen an unprecedented "explosion" in the availability of data from Earth observations obtained continuously from space and airborne sensors. This huge volume of Remote Sensing (RS) Data has been combined with high-performance computing (HPC) to make this data usable and available to the research community (Ma et al., 2015). As the sources and techniques for utilising this type of data proliferate, so do their applications in the social sciences (Dugoua et al., 2017) offering large-scale observations and measurements that would otherwise require prohibitively large financial resources to acquire. Concerning research infrastructures and data repositories, there are legal, ethical and data protection issues as well as technical issues due to the large volume and the data cleaning process. Furthermore, issues of accuracy of the data documentation and data file formats are raised to ensure the reusability and interoperability of the data.

The examples we present from the DANS repository are:

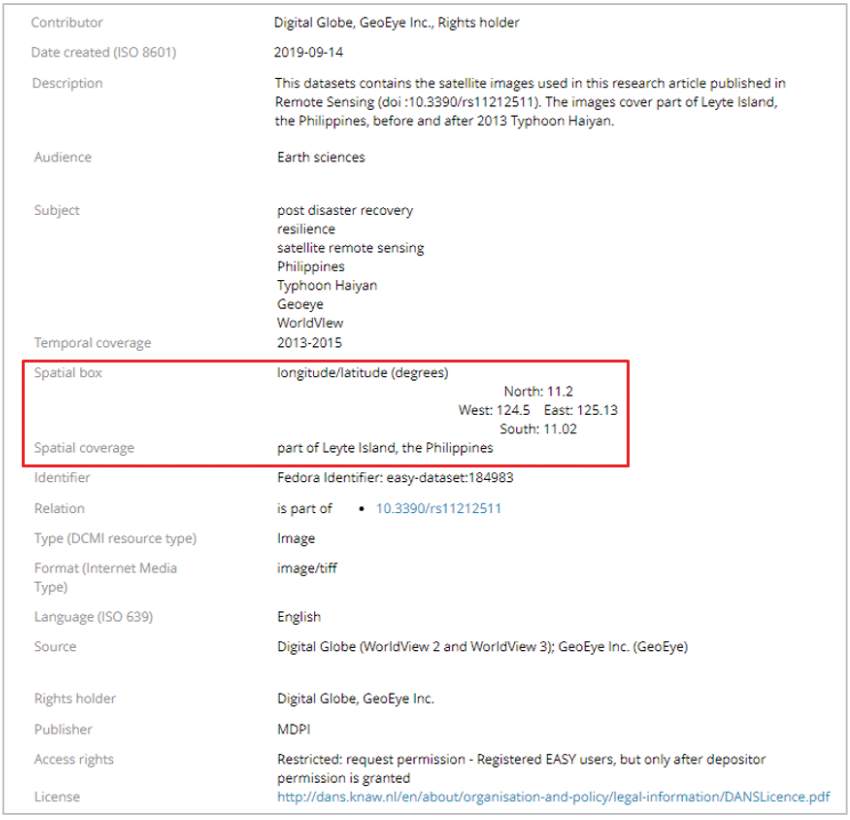
1. "*Long-term Assessment of Ecosystem Services at Ecological Restoration Sites Using LandSat Time Series*"[19]
2. "*Evaluating Resilience-Centered Development Interventions with Remote Sensing*"[20]

We observe that in both cases there is sufficient documentation of the data sources as well as the use of spatial boxes and spatial coverage in the geospatial metadata that accurately indicate the geographic area for which the data is available (Figure 4). However, we have different file formats. In the case of "*Evaluating Resilience-Centered Development Interventions with Remote Sensing*", image data in.tif format are available upon request, while in the case of "*Long-term Assessment of Ecosystem*

19. For access: https://doi.org/10.17026%2Fdans-zrc-hmz4
20. For access: https://doi.org/10.17026/dans-z99-7j8z

Services at Ecological Restoration Sites Using LandSat Time Series" there is unrestricted access to several different files such as:

1. Shapefiles
2. Textual documents such as software and shapefile descriptions in.pdf format and comma-separated values (.csv) files
3. Executable software files (.exe).



Contributor	Digital Globe, GeoEye Inc., Rights holder
Date created (ISO 8601)	2019-09-14
Description	This datasets contains the satellite images used in this research article published in Remote Sensing (doi :10.3390/rs11212511). The images cover part of Leyte Island, the Philippines, before and after 2013 Typhoon Haiyan.
Audience	Earth sciences
Subject	post disaster recovery resilience satellite remote sensing Philippines Typhoon Haiyan Geoeye WorldVIew
Temporal coverage	2013-2015
Spatial box	longitude/latitude (degrees) North: 11.2 West: 124.5 East: 125.13 South: 11.02
Spatial coverage	part of Leyte Island, the Philippines
Identifier	Fedora Identifier: easy-dataset:184983
Relation	is part of • 10.3390/rs11212511
Type (DCMI resource type)	Image
Format (Internet Media Type)	image/tiff
Language (ISO 639)	English
Source	Digital Globe (WorldView 2 and WorldView 3); GeoEye Inc. (GeoEye)
Rights holder	Digital Globe, GeoEye Inc.
Publisher	MDPI
Access rights	Restricted: request permission - Registered EASY users, but only after depositor permission is granted
License	http://dans.knaw.nl/en/about/organisation-and-policy/legal-information/DANSLicence.pdf

Figure 4: *Geospatial metadata in "Evaluating Resilience-Centered Development Interventions with Remote Sensing" in DANS repository*

HEALTH DATA

The recent outbreak of the COVID-19 pandemic came at a time when the principles of Open Science and FAIR data were fairly well established. The need for research and data in areas such as transmissibility, geographic spread, economic and psychological impact, risk factors for infection and even the circulation of fake news has given research infrastructures a central role in hosting and disseminating this data. Thus, we immediately saw the fast growth of already established repositories and the creation of new repositories and research projects related to COVID-19 daily cases, population mobility and transportation, research papers, health facilities, socioeconomic data, global and local news, social media, policy and regulations etc. (Hu et al., 2020; Kabir & Madria, 2020; Li et al., 2020; Reinhart et al., 2021). The need for data was so urgent that in several cases the fast-track procedures adopted violated some of the principles of Open Science (Besançon et al., 2021), often with debilitating consequences (Kadakia et al., 2021).

Health-related data is certainly not a new concept, but it was the first time that the imperative for well-timed and adequate dissemination became so urgent and covered more aspects than mere health-related administrative or statistical data. From a research infrastructure perspective, it became apparent that there will be an increasing demand for health-related data in the future, and efforts are being made to overcome the challenges posed by the unique character of such data.

We present two cases of health-related data hosted in the UKDA repository:

1. "*Administrative health data Brazil, 1996-2004*"[21]
2. "*Student Mental Health During Covid-19 Pandemic, 2020*"[22]

In both cases there is complete and standardised documentation, but we observe different levels of access for different types of data. In the first case of the administrative data for Brazil, the data is in Safeguarded access status, and users will need to log in to their account and follow the process set out by the UKDA repository to gain access (Figure 5).

In the case of the data from the "*Student Mental Health During Covid-19 Pandemic, 2020*" survey, as the data is derived from self-reported questions about the mental health of participants and is not the

21. For access: https://dx.doi.org/10.5255/UKDA-SN-852583
22. For access: https://dx.doi.org/10.5255/UKDA-SN-854720

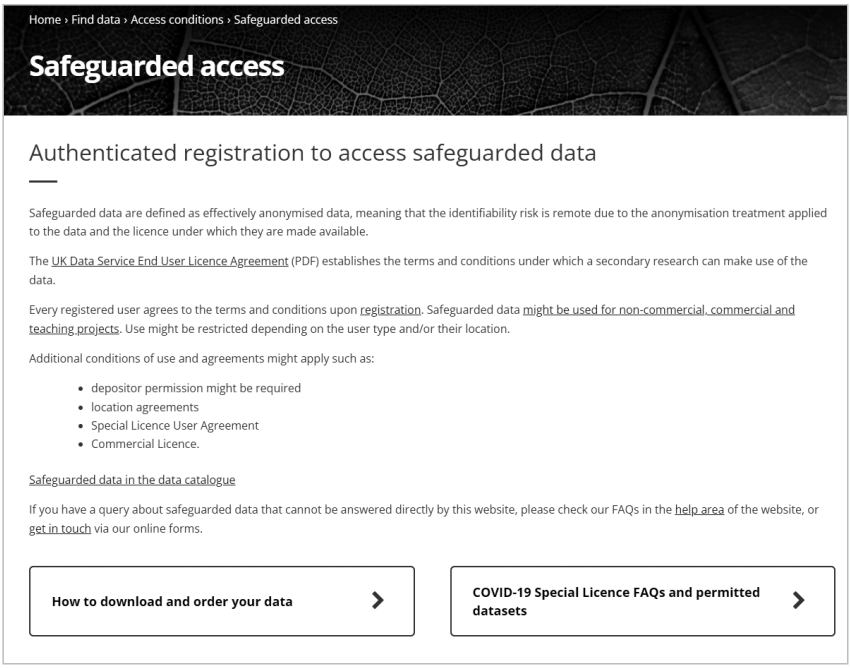


Figure 5: *Safeguarded data access requirements in UKDA repository*

result of observation or medical records, it is anonymised but available in a common interoperable format (.csv), as are the accompanying documents (codebook, questionnaire).

CONCLUSION AND CHALLENGES AHEAD

We reside within a complex network of connections, navigating through a myriad of networking approaches in our professional endeavours. Professionals involved in the world of data are encountering and will face many challenges and opportunities in the years to come. As researchers, we are called upon to make use of these new types of data, which contain the digital traces of our transactions and activities. We create "pictures" and instances of social organisation and contribute to the understanding of contemporary societies by searching, selecting and finding the appropriate sources in the deluge of data and information. Consequently, with much of the core social data now in textual form, which fundamentally changes how data are acquired and produced, researchers and scholars will need to come

to new agreements on what constitutes reliable and valid descriptions of the data, the categories used to organize those data, and the tools necessary to access, process, and structure those data. (Shah, Cappella & Neuman, 2015).

As data professionals, we are required to manage and maintain datasets generated from transaction traces consisting of personal activities as well as social and commercial aspects of the behaviour of individuals and groups. As such, the management of NDTs is a demanding process in terms of human resources, technical and professional/archival skills, as well as financial resources. Lack of user engagement and the securing of adequate resources in terms of staff capacity as well as financial resources also appear to be top challenges in a survey conducted among data professionals, regardless of the nature of the repositories (interdisciplinary, institutional), as reported in the context of an online survey designed among data repository managers (Khan et al., 2021).

The big assets in management practices of trusted Data repositories are the implementation of FAIR Data (Kondyli & Klironomos, 2022) as well as the provision of training services to researchers, guidance via the Data management Plan to design research as well as cooperation and exchange of know-how between peer institutions. To a large extent, Data Repositories are called upon to set up processes and broaden the range of collaborations so that they continue to be a suitable and certified space for hosting, documenting and synergizing the research process in the social sciences. Soon, the main challenges will be the ability of the research potential to work interdisciplinarily, taking into account the ethical, legal and technical issues of conducting research with new types of data. They will need to evolve and constantly improve services and tools for re-use. Re-use will also gradually allow the main actors of the research ecosystem (researchers, data managers and professionals, data institutions, funders, private and public institutions, etc.) to overcome barriers for the benefit of research and the public good. The first steps have already been taken.

REFERENCES

Ackland, R. (2013). *Web social science: Concepts, data and tools for social scientists in the Digital age*. SAGE Publications.

Athanasiou, S., Amiridis, V., Gavriilidou, M., Gerasopoulos, E., Dimopoulos, A., Kaklamani, G., Karagiannis, F., Klampanos, I., Kondyli, D., Koumantaros, K., Konstantopoulos, P., Lenaki, K., Likiardopoulos, A., Manola, N., Mitropoulou, D., Benardou, A., Boukos, N., Nousias, A., Ntaountaki, M., … Psomopoulos, F. (2020). *National Plan for Open Science*. Zenodo. https://doi.org/10.5281/zenodo.3908953

Barbier, G., Liu, H. (2011). Data mining in social media. In C. Aggarwal (Eds.), *Social Network Data Analytics* (pp. 327-352). Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-8462-3_12

Bauchner, H., Golub, R. M., & Fontanarosa, P. B. (2016). Data sharing. *JAMA*, *315*(12), p. 1238. https://doi.org/10.1001/jama.2016.2420

Besançon, L., Peiffer-Smadja, N., Segalas, C., Jiang, H., Masuzzo, P., Smout, C., Billy, E., Deforet, M., & Leyrat, C. (2021). Open science saves lives: Lessons from the covid-19 pandemic. *BMC Medical Research Methodology*, *21*(1). https://doi.org/10.1186/s12874-021-01304-y

Bastin, G. & Tubaro, P. (2018). Le moment big data des sciences sociales. Paris : Presses de Sciences Po. *Revue française de sociologie*. 2018/3 Vol. 59, 375-394. Αποθήκευση 21/9/2019. Url: https://www.cairn.info/revue-francaise-de-sociologie-2018-3-page-375.htm.

Bishop, L. (2017). Big data and data sharing: Ethical issues. UK Data Service, UK Data Archive, *7*.

Boté, J. J., & Termens, M. (2019). Reusing data technical and ethical challenges. *DESIDOC Journal of Library & Information Technology*, *39*(06), pp. 329–337. https://doi.org/10.14429/djlit.39.06.14807

Boyd, D., & Crawford, K. (2012). Critical questions for Big Data. *Information, Communication & Society*, *15*(5), pp. 662–679. https://doi.org/10.1080/1369118x.2012.678878

Breuer, J., Bishop, L., & Kinder-Kurlanda, K. (2020a). The practical and ethical challenges in acquiring and Sharing Digital Trace Data: Negotiating public-private partnerships. *New Media & Society*, *22*(11), pp. 2058–2080. https://doi.org/10.1177/1461444820924622

Chou, M. H. (2014). The evolution of the European Research Area as an idea in European integration. *Building the Knowledge Economy in Europe*, pp. 27–50. https://doi.org/10.4337/9781782545293.00007

Conrad, F. G., Gagnon-Bartsch, J. A., Ferg, R. A., Schober, M. F., Pasek, J., & Hou, E. (2019). Social media as an alternative to surveys of opinions about the economy. *Social Science Computer Review*, 089443931987569. https://doi.org/10.1177/0894439319875692

Desrosières, A. (2005). Décrire l'État ou explorer la société: les deux sources de la statistique publique. *Genèses* 2005/1 (no 58), pp. 4-27. [In French]

Digital Science (2017). The State of Open Data Report 2017. Digital Science. https://doi.org/10.6084/m9.figshare.5481187

Dugoua, E., Kennedy, R., & Urpelainen, J. (2018). Satellite data for the Social Sciences: Measuring rural electrification with night-time lights. *International Journal of Remote Sensing*, *39*(9), pp. 2690–2701. https://doi.org/10.1080/01431161.2017.1420936

ESFRI (2006). European Roadmap for Research Infrastructures. 2006 Report. Office for Official Publications of the European Communities. Belgium. https://www.esfri.eu/sites/default/files/esfri_roadmap_2006_en.pdf

ESFRI (2008). European Roadmap for Research Infrastructures. 2006 Report. Office for Official Publications of the European Communities. Belgium. https://www.esfri.eu/sites/default/files/esfri_roadmap_update_2008.pdf

Giglietto, F., & Rossi, L. (2012). Ethics and Interdisciplinarity in computational social science. *Methodological Innovations Online*, *7*(1), pp. 25–36. https://doi.org/10.4256/mio.2012.003

Harari, G. M., Müller, S. R., Aung, M. S., & Rentfrow, P. J. (2017). Smartphone sensing methods for studying behavior in everyday life. *Current Opinion in Behavioral Sciences*, *18*, pp. 83–90. https://doi.org/10.1016/j.cobeha.2017.07.018

Harford, T. (2014), Big data: A big mistake?. *Significance*, *11*, pp. 14-19. https://doi.org/10.1111/j.1740-9713.2014.00778.x

Hemphill, L., Pienta, A., Lafia, S., Akmon, D., & Bleckley, D. (2022). How do properties of data, their curation, and their funding relate to reuse?. *Journal of the Association for Information Science and Technology, 73*(10), 1432–1444. https://doi.org/10.1002/asi.24646

Hox, J. J., & Boeije, H. R. (2005). Data Collection, Primary vs. Secondary. In *Encyclopedia of Social Measurement* (pp. 593-599). Elsevier, Amsterdam. https://doi.org/10.1016/B0-12-369398-5/00041-4

Hu, T., Guan, W. W., Zhu, X., Shao, Y., Liu, L., Du, J., Liu, H., Zhou, H., Wang, J., She, B., Zhang, L., Li, Z., Wang, P., Tang, Y., Hou, R., Li, Y., Sha, D., Yang, Y., Lewis, B., … Bao, S. (2020). Building an open resources repository for covid-19 research. *Data and Information Management*, *4*(3), pp. 130–147. https://doi.org/10.2478/dim-2020-0012

Kabir, M., & Madria, S. (2020). CoronaVis: a real-time COVID-19 tweets data analyzer and data repository. *arXiv preprint* arXiv:2004.13932.

Kadakia, K. T., Beckman, A. L., Ross, J. S., & Krumholz, H. M. (2021). Leveraging open science to accelerate research. *New England Journal of Medicine*, *384*(17). https://doi.org/10.1056/nejmp2034518

Karpf, D. (2012). Social Science Research Methods in internet time. *Information, Communication & Society*, *15*(5), pp. 639–661. https://doi.org/10.1080/1369118x.2012.665468

Khan, N., Thelwall, M. & Kousha, K. (2022). Are data repositories fettered? A survey of current practices, challenges and future technologies. *Online Information Review, 46*(3), pp. 483-502. https://doi.org/10.1108/OIR-04-2021-0204

King, G. (2011). Ensuring the data-rich future of the Social Sciences. *Science*, *331*(6018), pp. 719-721. https://doi.org/10.1126/science.1197872

Kleiner, B., Kondyli, D., Klironomos, N., Bishop, L., Vavra, M. & Cizek, T. (2022). D14 Overview and summary of existing outputs (inside and outside of CESSDA) on NDTs. CESSDA.

Kondyli, D. & Klironomos, N. (2022). FAIR Data: Opportunities and challenges for research infrastructures and research communities. In J. Kallas et al. (Eds.), Development of Infrastructures for Data Production and Management in the Social Sciences. [In Greek]. https://doi.org/10.17903/CV09INFRA

Kondyli, D. & Linardis A. (2021). New data types, new roles for research infrastructures?. In N. Nagopoulos (Ed.), *Social Sciences today. Dilemmas and perspectives beyond the crisis*. Proceedings of the 2nd conference of the School of Social Sciences, University of the Aegean. [In Greek]

Kondyli, D., Nisiotis, C. S., & Klironomos, N. (2024). Data reusability for migration research: A use case from SoDaNet data repository. *Frontiers in Human Dynamics, 5*. https://doi.org/10.3389/fhumd.2023.1310420

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a

research tool for the Social Sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, *70*(6), pp. 543–556. https://doi.org/10.1037/a0039210

Lagoze, C., Block, W. C., Williams, J., Abowd, J., & Vilhuber, L. (2013). Data Management of Confidential Data. *International Journal of Digital Curation, 8*(1), pp. 265–278. https://doi.org/10.2218/ijdc.v8i1.259

Lazer, D., & Radford, J. (2017). Data ex machina: Introduction to big data. *Annual Review of Sociology*, *43*(1), pp. 19–39. https://doi.org/10.1146/annurev-soc-060116-053457

Li, Y., Jiang, B., Shu, K., & Liu, H. (2020). Toward a multilingual and multimodal data repository for covid-19 disinformation. *2020 IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/bigdata50022.2020.9378472

Linardis, A., Alexandris, K. & Klironomos, N. (2022). The new SoDaNet Data Catalogue. The transition from Nesstar to Dataverse. In J. Kallas et al. (Eds.), *Development of Infrastructures for Data Production and Management in the Social Sciences* (pp. 147-183). [In Greek]. https://doi.org/10.17903/CV06INFRA

Ma, Y., Wu, H., Wang, L., Huang, B., Ranjan, R., Zomaya, A., & Jie, W. (2015). Remote sensing big data computing: Challenges and opportunities. *Future Generation Computer Systems*, *51*, pp. 47–60. https://doi.org/10.1016/j.future.2014.10.029

Mannheimer, S., & Hull, E. A. (2018). Sharing selves: Developing an ethical framework for curating Social Media Data. *International Journal of Digital Curation*, *12*(2), pp. 196–209. https://doi.org/10.2218/ijdc.v12i2.518

Metzler, K., Kim, D. A., Allum, N., & Denman, A. (2016). Who is doing computational social science? Trends in big data research (White paper). SAGE Publishing. https://doi.org/10.4135/wp160926

OECD (2013). *New data types for understanding the human condition*. https://www.oecd.org/sti/inno/new-data-for-understanding-the-human-condition.pdf

OECD (2016). Research ethics and new forms of data for social and Economic Research. OECD Science, Technology and Industry Policy Papers. https://doi.org/10.1787/5jln7vnpxs32-en

Politou, E., Alepis, E., Virvou, M., & Patsakis, C. (2021). Conclusions. *Privacy and Data Protection Challenges in the Distributed Era*, pp. 181–185. https://doi.org/10.1007/978-3-030-85443-0_11

Reinhart, A., Brooks, L., Jahja, M., Rumack, A., Tang, J., Agrawal, S., Al Saeed, W., Arnold, T., Basu, A., Bien, J., Cabrera, Á. A., Chin, A., Chua, E. J., Clark, B., Colquhoun, S., DeFries, N., Farrow, D. C., Forlizzi, J., Grabman, J., … Tibshirani, R. J. (2021). An open repository of real-time covid-19 indicators. *Proceedings of the National Academy of Sciences*, *118*(51). https://doi.org/10.1073/pnas.2111452118

Resnik, D. B., & Elliott, K. C. (2015). The ethical challenges of socially responsible science. *Accountability in Research*, *23*(1), pp. 31–46. https://doi.org/10.1080/08989621.2014.1002608

Ruppert, E., Law, J., & Savage, M. (2013). Reassembling social science methods: The Challenge of Digital Devices. *Theory, Culture & Society*, *30*(4), pp. 22–46. https://doi.org/10.1177/0263276413484941

Salah, A. A., Canca, C., & Erman, B. (2022). Ethical and legal concerns on data science for large scale human mobility. In A. A., Salah et al. (Eds.), *Data Science for Migration and Mobility*. Proceedings of the British Academy. https://webspace.science.uu.nl/~salah006/salah22legal.pdf

Savage, M. (2016). The use of big data in the analysis of inequality. In ISSC, IDS and UNESCO,

Challenging Inequalities: Pathways to a Just World, World Social Science Report. UNESCO Publishing http://en.unesco.org/wssr2016

Sawchuk, S. L., & Khair, S. (2021). Computational reproducibility: A practical framework for data curators. *Journal of eScience Librarianship, 10*(3). https://doi.org/10.7191/jeslib.2021.1206

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big Data, digital media, and Computational Social Science. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), pp. 6–13. https://doi.org/10.1177/0002716215572084

Silber, H., Breuer, J., Beuthner, C., Gummer, T., Keusch, F., Siegers, P., Stier, S., & Weiß, B. (2022). Linking surveys and digital trace data: Insights from two studies on determinants of data sharing behaviour. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(Supplement_2). https://doi.org/10.1111/rssa.12954

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2019). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, *38*(5), pp. 503–516. https://doi.org/10.1177/0894439319843669

Stuart, D., Baynes, G., Hrynaszkiewicz, I., Allin, K., Penny, D., Lucraft, M., & Astell, M. (2018). Whitepaper: Practical challenges for researchers in data sharing. https://doi.org/10.6084/m9.figshare.5975011.v1

Thanos, C. (2017). Research data reusability: Conceptual Foundations, barriers and Enabling Technologies. *Publications*, *5*(1), 2. https://doi.org/10.3390/publications5010002

Ulnicane, I. (2019). Broadening aims and building support in science, technology and innovation policy: The case of the European Research Area. *Journal of Contemporary European Research*, *11*(1), pp. 31-49.

Uzwyshyn, R. (2016) Online Research Data Repositories: The What, When, Why and How. *Computers in Libraries, 36*(3), pp. 18-21. https://digital.library.txstate.edu/handle/10877/7597

Zhou, B., Pei, J., & Luk, W. (2008). A brief survey on anonymization techniques for privacy preserving publishing of Social Network Data. *ACM SIGKDD Explorations Newsletter*, *10*(2), pp. 12–22. https://doi.org/10.1145/1540276.1540279