seriss SYNERGIES FOR EUROPE'S RESEARCH INFRASTRUCTURES IN THE SOCIAL SCIENCES

Deliverable Number: D6.5

Deliverable Title: Guidelines on linking administrative data with survey data for research

Work Package: 6: New forms of data: legal, ethical and quality matters

Deliverable type: Report

Dissemination status: Public

Submitted by: NIDI

Authors: Straume, Ø. (CESSDA, NSD), L'Hours, H. (CESSDA, UKDA), Emery, T. (NIDI, GGP), Rød, L.M. (CESSDA, NSD), Høgetveit Myhren, M. (CESSDA, NSD), Bishop E. (CESSDA, GESIS), Hagen, S. (CESSDA, NSD), Grøndahl, P.E. (CESSDA, NSD)

Date submitted: August 2019

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 654221.

www.seriss.eu @SERISS_EU

SERISS (Synergies for Europe's Research Infrastructures in the Social Sciences) aims to exploit synergies, foster collaboration and develop shared standards between Europe's social science infrastructures in order to better equip these infrastructures to play a major role in addressing Europe's grand societal challenges and ensure that European policymaking is built on a solid base of the highest-quality socio-economic evidence.

The four year project (2015-19) is a collaboration between the three leading European Research Infrastructures in the social sciences – the European Social Survey (ESS ERIC), the Survey for Health Aging and Retirement in Europe (SHARE ERIC) and the Consortium of European Social Science Data Archives (CESSDA AS) – and organisations representing the Generations and Gender Programme (GGP), European Values Study (EVS) and the WageIndicator Survey.

Work focuses on three key areas: Addressing key challenges for cross-national data collection, breaking down barriers between social science infrastructures and embracing the future of the social sciences.

Please cite this deliverable as: Straume, Ø., L'Hours, H., Emery, T., Rød, L.M., Høgetveit Myhren, M, Bishop E, Hagen, S. Grøndahl, P.E, (2019) Guidelines on linking administrative data with survey data for research. Deliverable 6.5 of the SERISS project funded under the *European Union's Horizon 2020 research and innovation programme* GA No: 654221. Available at: wwwww.seriss.eu/resources/deliverables www.seriss.eu/resources/deliverables

Content

1 Introduction

Administrative data are increasingly in demand for research purposes. The increased availability of administrative data as a basis for survey sampling methodologies increases the scope and capacity for linking administrative data and survey data for research purposes. The potential for such linkage is enhanced by the bridging of scientific and civil service data infrastructures and the deeper integration of science and the policy making process. By linking data, policy makers can validate their data with scientifically driven measures and social scientists can better engage in questions with direct relevance for policy makers.

This potential is not easily realized. The data infrastructures of administrative data sources and social surveys are generally governed by a variety of laws, different ethical practices and different priorities and purposes. Operating across these differences poses several challenges.

This report addresses the legal requirements and ethical issues in linking administrative data and survey data and aims to provide guidelines on use of administrative data in survey research.

Work Package 6 of the SERISS project addresses the major legal and ethical challenges facing cross-national social science research which relies on access to large-scale data on an individual level. The focus is on social surveys and the use of new data types in a social survey context in particular, including biomarker, social media data and administrative data.

The main focus of Task 6.4 was the legal requirements and ethical challenges that may come about when survey data are linked with administrative data sources. Administrative data is defined as data that is collected on a non-voluntary basis as part of government activities. The task addresses the issues that need to be taken to meet these challenges in order to increase and improve the research use of these data sources[1].

The main purpose of this report is to provide guidance to the legal and ethical issues in, and to overcome possible barriers to, research use of administrative data. A set of guidelines is crucial to increase opportunities for linkage within the boundaries of the GDPR. As such this deliverable is aimed at researchers and infrastructures managing new forms of data.

These guidelines are structured as Questions and Answers that have been selected with legal and ethical issues in mind. The guidelines aim to work as advice to researchers and research infrastructures looking to engage in administrative data linkage. The questions are structured around the research data lifecycle, covering research design and data collection, data processing and archiving.

[1] European Commission, Directorate-General for Research and Innovation (2016: 45).

2 Questions and Answers

Research Design & Data Collection

When do I need to think about the legality and ethics of administrative data linkage?

If you intend to link your research data to administrative data sources' then it is optimal and often necessary to consider this at the very outset of your project. From a legal, ethical and technical perspective, the linkage to administrative data must be incorporated into the research design and it is difficult to add on to an existing research project post-hoc. From a legal perspective, linking to administrative data can often depends on the research project aims and scope aligning with the legally mandated objectives of the administrative data provider. To link with such administrative data, it can be helpful to ensure that the research question under investigation is in alignment with the aims and scope of the administrative data provider.

It can also be necessary to acquire consent for linkage. The legal requirements regarding consent vary considerably depending on the context. In addition, ethical codes can require explicit and informed consent for administrative data linkage regardless of the legal requirements. From a technical perspective, administrative data linkage needs to be considered at the outset to ensure that the required identifiers and linkage variables are collected. Furthermore, in many areas of research the integration of administrative data linkage at the design stage can help the researcher leverage the administrative data to improve the overall research design such as by using the administrative data records as a sampling frame.

Who should I contact about the legal and ethical aspects of linkage?

Some administrative data providers have support for research within their general scope of work. These organizations commonly have service desks who support administrative data linkage. This is the case for statistical offices whose general work is to support scientific research as well as statistical processing for government purposes. To further facilitate these service desks, many research councils have financed the development of support services for administrative data linking that are independent of administrative data providers. For example, the UK Data Archive, DANS/ODISSEI in the Netherlands and NSD in Norway have all provided such services in some form. National Data archives are generally the host or coordinators of such initiatives and can be a useful source of information and guidance in approaching administrative data providers and advising on the whole data lifecycle.

In some instances, researchers may seek to link with administrative data sources that are not controlled by organizations which have support for research as part of their mandate. For example, ministries for education and health are less likely to have a dedicated service desk for the support of scientific research. In such cases, contacting the data protection officer of such organizations can be a useful first step in understanding what type of data the organization holds and under what basis it is processed. These are important pieces of information when formulating a research question and understanding the potential for data linkage. The contact information for the data protection officer for any such organization is publicly available as required by the GDPR.

When do I need consent for data linkage?

Under the General Data Protection Regulation (GDPR), personal data can be processed for several different reasons. For social research, there are two principle bases on which personal data may be processed; because consent has been acquired or because it is for scientific or historical research purposes or statistical purposes. Administrative data is generally designed to be processed because it is in the public interest to do so (Article 6, Article 9 and Recitals (51) to (56) of the GDPR). That is to say the administrative purposes of doing so are in the public interest. This does not require consent and the collection of such data can be mandated by law so long as specific protections are provided for (Article 89(1)). By contrast, survey data is predominantly processed based on the respondents' consent. When linking administrative data and survey data, processing is regularly on two different grounds.

Furthermore, the specific national derogations on scientific usage and usage in the public interest mean that the interactions between these two grounds is specific to the national context. The interests of the data subject, the protections required and the data processing procedures these necessitate can therefore be exceptionally complex and potentially contradictory. It is accepted good practice within the context of the GDPR to process data on a single ground to avoid such issues. It is therefore advisable that researchers utilize consent as a basis for processing where possible. Processing data for research on a public interest basis is possible but requires close cooperation with the administrative data provider to ensure that the research activities are in line with the legal basis upon which they process data and that the administrative data provider is data controller within the research project itself.

What about using administrative data for sampling?

Administrative data can provide high quality sampling frames that are not otherwise publicly accessible or covered by the primary scope of activities for the administrative data provider. It is impossible to draw a sample from administrative data on a consent basis because it is prior to contact with respondents. The legal basis of drawing a sample is therefore dependent on national derogations and the legal basis of processing for the administrative data provider.

The sample can generally be drawn and processed on the basis of scientific or historical research purposes that are not considered incompatible with the data's original purpose (Article 6(1e); Article 89(1)). In such instances it is necessary that the controller for the administrative data provide data subjects with information as detailed by Article 14 of the GDPR. Informing subjects is not obligatory if it entails, 'disproportionate effort' but in cases where the purpose of drawing a sample is to subsequently contact individuals thus could not be justified.

What information is required for data subjects when processing personal data?

The GDPR lists the categories of information that must be provided to a data subject in relation to the processing of their personal data where it is collected from the data subject (Article 13) or obtained from another source (Article 14). For an overview of the required information see the Article 29 working party guidelines on transparency, pages 35-40.[2] When administrative data is used as a basis of sampling, both Article 13 and Article 14 apply, and invitation letters and contact sheets must detail the information stipulated in both. The information required is

(a) the identity and the contact details of the controller and, where applicable, of the controller's representative;

(b) the contact details of the data protection officer, where applicable;

(c) the purposes of the processing for which the personal data are intended as well as the legal basis for the processing;

(d) the categories of personal data concerned;

(e) the recipients or categories of recipients of the personal data, if any;

(f) the period for which the personal data will be stored, or if that is not possible, the criteria used to determine that period;

(g) where the processing is based on point (f) of Article 6(1), the legitimate interests pursued by the controller or by a third party;

(h) the existence of the right to request from the controller access to and rectification or erasure of personal data or restriction of processing concerning the data subject and to object to processing as well as the right to data portability;

(i) the right to lodge a complaint with a supervisory authority;

(j) from which source the personal data originate, and if applicable, whether it came from publicly accessible sources;

(k) the existence of automated decision-making, including profiling.

There are also potential national derogations that could apply, and the legal basis of processing can also influence the type of information that is included such as the measures taken to ensure that the legal basis is adhered to.

What is granular consent and what does it look like?

The best way to achieve informed consent for data sharing is to identify and explain the possible future uses of their data, and to offer the participants the option to consent on a granular level where relevant and possible. The guidelines on Consent[3] from the Article 29 Working Party state that "…when data processing is done in pursuit of several purposes, the solution to comply with the conditions for valid consent lies in granularity, i.e. the separation of these purposes and obtaining consent for each purpose." For administrative data linkage one approach could be to allow the participants the choice of what the data can be processed for. This could be based on explicit research project aims, the types of data that can be linked or the types of researchers who access the data. It is important to note however, that it is then the responsibility of the data controller to oversee and monitor

[2] http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=622227
[3] https://iapp.org/resources/article/wp29-guidelines-on-consent/

whatever granularity that is provided. It is therefore in the interest of the controller that such granularity is well defined, explicit and clear to facilitate this.

It is worth noting that administrative data is not collected on a consent basis and is often done so passively without explicit agreement by the data subject (for example, tax records). It is therefore advisable to explicitly state the precise data that are to be linked so as to adhere to Article 7 of the GDPR which states that the information provided should be in an intelligible and easily accessible form, using clear and plain language. This could include clear examples of the data collected and the way in which it is collected. A granular consent could therefore look something like this:

Please tick the appropriate boxes	yes	no
I agree to take part in the survey described in the information letter.	□	□
I agree to allow the project to collect my data from [administrative data provider] including [administrative data records]	□	□
I allow the project to link my data from [administrative data provider] to my survey responses.	□	□

Data Processing

When do I process personal data?

The GDPR defines 'personal data' as any information that can be used to identify a person ('data subject'), directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity (Article 4 (1)).

When designing research with an administrative data linkage component, it is worth noting that identifiers for individuals and households that are used by the government have been included in the definition of personal data in the new law.

Article 4 in the GDPR gives us the following definition of processing:

‘processing’ means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction;

For administrative data this means that social insurance numbers, tax codes, driving license number and other identification numbers are all included under the definition of personal data. So long as an administrative linkage key exists all data is considered personal data for processing purposes.

What can I do if I have consent from respondents to link data?

The GDPR provides for the ‘Right to Data Portability’ (Article 20). When linking to social media sources, this right can be utilized such that consent is sufficient basis for data processing and social media platforms are required to provide individuals with machine readable data that they can then make available for linkage. However, data controllers who are processing data on public interest grounds are not subject to the Right of Data Portability. Administrative data is almost exclusively processed on a public interest basis and is therefore not subject to this right. Administrative data providers are therefore under no obligation to enable the linkage of social research data and administrative data even in instances where explicit consent is given by the data subject. This means that it is advisable to attain the cooperation of the relevant administrative data providers prior to data collection and the acquisition of consent to ensure that it would permit linkage on the legal basis under which the administrative data provider processes data. Administrative data providers are under no obligation to cooperate post-hoc.

It is also important to note that whilst linking data based on consent can allow for data be processed under terms defined within the informed consent (i.e. outside of a secure or restricted access environment), linking on a consent basis can constrain the types of information that can be linked. When using consent, the data to which one links is limited to only the data subject that has provided consent. For example, it would not be possible to link the income of a data subjects’ spouse or parents, even though these are often a key variable of interest in social research. Much of the scientific potential in administrative data linkage lies in the ability to link individuals with broader social networks and processes but many of these are not possible when linking on the basis of consent. The implication of this is that to conduct such analysis the data controller within such scientific projects must be the administrative data provider and such analysis must be done within a secure environment. Researchers need to think carefully about what data they want to link and how they want to analyze the data before committing to a basis for linkage.

When is administrative data anonymous?

Administrative data is only anonymized when the linkage key to all personal data is destroyed and reasonable precautions have been undertaken to prevent data disclosure. The risk of data disclosure in these circumstances is highly dependent on the institutional and legal context of the administrative data provider and the nature of the data linked. Certain administrative data can be redacted or masked in such a way as to mitigate disclosure risks, such as the collapsing of income data into percentiles or the use of established masking techniques. However, the primary value in administrative data linkage lies precisely in the fact that it provides personal information on data subjects. Administrative data is therefore seldom anonymous and is generally subject to the GDPR.

The responsibility of ensuring that disclosure risks have been sufficiently mitigated using masking or encryption techniques lies with the data controller. If the data is sufficiently anonymized however, the data then falls outside of the GDPR as it is no longer personal data. Specific recommendations on pseudonymization and general anonymization procedures are provided by the Article 29 working party and can provide guidance on how best to achieve this. Nevertheless, even once a dataset has been deemed by the controller to be anonymized a proper reporting mechanism for disclosure risks needs to be maintained and information of the data source and contact details of the data protection officer should always be included in the metadata of anonymized data derived from administrative data.

This way, if a breach is identified, the data controller can be notified, and they can fulfil their obligations under the GDPR.

How do I analyze linked Administrative Data?

If the data has been anonymized, it can be analyzed in the same way as any other social science dataset. For example, the Generations and Gender Programme includes linked administrative data from the Netherlands, Sweden and Norway as the data included underwent a disclosure risk assessment. These data therefore are archived and distributed alongside the general survey data. Administrative data which is not anonymized or has not been collected based on consent that allows for distribution cannot be distributed in such a way. This is because the items of interest from the linked administrative data are by nature a disclosure risk. Therefore, most analysis of linked administrative data occurs in secure, controlled environments. Especially when the basis of processing is for scientific research, the data controller must be able to ensure that analysis is limited to this and therefore various technical infrastructures have been developed which allowed controlled access to administrative data.

For example, data from SHARE wave 1-5 in Denmark has been linked to the administrative records from the Danish Health Authority and Statistics Denmark without consent from respondents. The analysis of the linked data must therefore be conducted via a secure remote access facility from a Danish research institution. This can limit the types of analysis as such environments often do not have internet access, limiting the import of new statistical packages and the updating of statistical software.

The specific parameters of these secure environments vary considerably. Some require onsite access and physical checks and oversite through a ‘safe room’. Others allow for remote access into a secure online environment. These environments are however all highly regulated. They also make the researcher dependent on the cooperation of the administrative data provider for access approval, and the secure environment provider for their analytical tools. The administrative data provider is under no obligation to provide access. Access can be limited, slow, constrained in terms of time and resources and subject to the cooperation of the administrative data provider. This is important for researchers to account for in project planning and risk management within their own work.

How can I document my workflow in a secure environment?

Administrative data is not transparent by design. It rarely comes with publicly available metadata and the data production processes are often opaque. This makes documentation of data processing with linked administrative data oagreater imperative even than in other forms of scientific analysis. However, data processing within a secure environment is difficult. This is because any output from the environment is subject to disclosure checks that are resource intensive. Adding additional metadata and workflow documentation to this is therefore not regularly practiced.

However, properly documenting the data analysis process with linked administrative data is essential for replicability, accountability and transparency[4]. When engaging in administrative data linkage it is therefore important to develop a coherent workflow documentation plan at the outset before access is granted. As part of access procedures, researchers are often required to stipulate what software they require access to in the secure environment and it is important that researchers include documentation tools such as Jupyter Notebooks in this list of software to enable them to thoroughly document their work. Even if the researcher doesn't use these tools in their regular work, they are especially important in administrative data linkage. Normally if a researcher is asked to provide additional information about analysis they have published, they can re-access syntax or data files they used during the analysis. This is often not as easy when the analysis was in a secure environment and researchers need to plan the documentation of their workflows accordingly.

Furthermore, when researchers have accessed the data through a secure environment it is good practice to document the access procedure and seek clarification from the data controller about any plans for sustainability of the data access process that exist. This will facilitate replication of any analysis conducted within the secure environment even in instances where the researcher is not themselves the data controller.

Archiving

What metadata should be stored for linked administrative data?

Creating and providing access to metadata is of increased importance for linked administrative data given that the data itself can often not be made readily available. To ensure transparency and replicability it is therefore essential that the administrative data that is analyzed is fully documented as part of the research process and that enough time and resources are incorporated into the research design to cover this. To make matters more challenging for researchers, administrative data is rarely fully documented for scientific purposes given that it is primarily designed for administration[5].

There are also no specific standard metadata formats for administrative data for use by social researchers. Metadata standards for the social sciences more generally, such as DDI, can be used to document administrative data sources but doing so is not easy. There are resources available to assist researchers in documentation such as CESSDA's training suite (https://www.cessda.eu/Training/). However, administrative data documentation is particularly difficult.

A researcher using linked administrative data will generally find themselves in a position where their obligation to document data is high, relative to secondary data sources such as surveys, and their knowledge of the data is relatively low or incomplete, relative to a project in which they are collecting their own data. It is therefore highly recommended that a

[4] Playford, Christopher J., Vernon Gayle, Roxanne Connelly, and Alasdair JG Gray. "Administrative social science data: The challenge of reproducible research." Big Data & Society 3, no. 2 (2016): 2053951716684143.

[5] Connelly, Roxanne, Christopher J. Playford, Vernon Gayle, and Chris Dibben. "The role of administrative data in the big data revolution in social science research." Social Science Research 59 (2016): 1-12.

researcher or research infrastructure looking to utilize administrative data involve a national data archive and see what kind of support can be offered in the documentation of their work.

How can I archive administrative data?

"Open access" principles for scientific information also apply to data. These principles have been formulated at an international level - for example, the OECD (2007, Principles and Guidelines for Access to Research Data from Public Funding). Making research data available for reuse is in line with requirements and/or recommendations from funding agencies and publishers that calls for a more open science[6]. Research Councils, funders, journals and publishers increasingly require data to be shared or deposited, or encourage researchers to share data. This is also compliant with the FAIR principles for reuse of research data.

However, administrative data can be rarely archived in a way that makes it accessible as administrative data providers are generally required to maintain full control over their data and cannot delegate responsibility for access to a data archive. Nevertheless, archiving the data at a data archive without the provision of access to other researchers is still exceptionally valuable. Even the storage of only the available metadata still satisfy the FAIR principles, especially if they clearly document and update the procedure and contact points for gaining access to the data. Here it is important to note the difference between FAIR and open data[7].

It should not be presumed that the administrative data provider is an adequate long-term archive. Government ministries and departments are not permanent bodies, and archiving processes are not always central in software or hardware updates. It is therefore necessary that archiving involves a trusted repository of sorts.

Is there some infrastructure for archiving administrative data?

No, there is currently no exclusive infrastructure for archiving linked administrative data. Existing administrative data sources that have been archived are generally done through the national data archive. Many national governments maintain internal government archives and data infrastructures, but these are generally not suitable for scientific research purposes. Statistical offices often maintain a data catalogue of administrative data and the architecture of government data resources, but these are not scientific repositories that are compliant with the FAIR principles for research data. They are designed for the wider use of data both within and outside government. One of the benefits of using a trusted data depository is that they offer more than a storage facility, since most of them also curate research data. This is to maintain the usability, understandability and authenticity of the data also in the future[8]. This is however and area which is rapidly developing and it could be that the situation has changed since the time of writing.

[6] Weller and Kinder-Kurlanda (2016)

[7] https://www.go-fair.org/faq/ask-question-difference-fair-data-open-data/

[8] https://www.cessda.eu/Research-Infrastructure/Training/Expert-Tour-Guide-on-Data-Management/6.-Archive-Publish/Publishing-with-CESSDA-archives

How do I prepare my data for sharing?

There is no clear guidance on how to archive and share administrative data, and the various approaches differ for example in terms of size and range of datasets (samples of data vs. population data).

First, you have to make sure that the data can be shared in accordance with legal and ethical frameworks. If personal data is to be archived the legal relationship with the data owner must be clear. Full written authorization from the administrative data provider must be provided in full as well as a clear communication of the responsibilities of all data users to report any potential data disclosure issues.

Research documentation

When depositing data within a data depository data should be clearly documented on study and data level, to make sure that the data can be understandable also in the future. Due to the novelty of administrative data, standards and methods are not firmly established across disciplines[9]. Recommendations on how to document and format your data can be found in data repositories that usually follow a metadata standard.

General information about the study such as study title, data type and research procedures should be provided. In the case of administrative data, documentation can also include explanations about the source of the data and the specific process by which it was collected. Detailed documentation of the provision of code and syntax can allow everyone to check how collection, cleaning and analysis have been performed. Other relevant documentation may include technical reports, workflows and description of analyzes. One will need to know exactly how the research was performed and what have been done as well as full data provenance of the administrative data and any quality assessment work undertaken on it.

Plan and prepare

Researchers looking to conduct administrative data linkage should make a data management plan to make sure that the checks such as consent and data ownership are in order. A data management plan is a tool where you can describe strategies and systematize the data management throughout the research life cycle. It can also save time and resources after the project have ended. National data archives provide templates and tools to facilitate the development of a data management plan.

Glossary

Administrative data

Administrative data refers to the data generated when individuals or organizations engage with public or private bodies for reasons other than scientific research. It often includes personally identifiable data.[10]

Administrative data sources

Administrative data sources are the public or private bodies who collate administrative data on individuals or organizations.

Anonymous data

data that cannot identify individuals in the data set, neither directly through name or social security number, indirectly through background variables, nor through a list of names or through an encryption formula and code/scrambling key.[11]

Consent

consent’ of the data subject means any freely given, specific, informed and unambiguous indication of the data subject’s wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her.[12]

Data access

the activity by which a researcher is given access to data.[13]

Data controller

‘controller’ means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data.[14]

Data curation

The activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purposes, and available for discovery and re-use. A more

[10] ESOMAR Guideline on Social Media Research (2011: 1-2)

[11] http://www.nsd.uib.no/personvernombud/en/help/vocabulary.html

[12] GDPR Article 4

[13] OECD 2016: 41

[14] GDPR Article 4

formal definition is: Digital curation is all about maintaining and adding value to a trusted body of digital information for future and current use: specifically, the active management and appraisal of data over the entire life cycle.[15]

Direct identifier

A person will be directly identifiable through name, social security number or other uniquely personal characteristics.

Data processor

Processor means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller.[16]

FAIR principles

A set of guiding principles to make data Findable, Accessible, Interoperable, and Reusable.[17]

General Data Protection Regulation

The GDPR (Regulation (EU) 2016/679) is a regulation by which the European Parliament, the European Council and the European Commission intend to strengthen and unify data protection for individuals within the European Union (EU). It came into force in May 2018.

Indirect identifiers

A person will be indirectly identifiable if it is possible to recognize the person through background information such as place of residence or institutional affiliation, combined with data on age, gender, occupation, diagnosis, etc.[18]

Long-term preservation

In data management, data preservation is the process of maintaining access to data so that it can still be found, understood and used in the future.[19]

[15] OECD 2016: 41

[16] GDPR Article 4

[17] https://www.force11.org/group/fairgroup/fairprinciples

[18] http://www.nsd.uib.no/personvernombud/en/help/vocabulary.html

[19] https://www2.le.ac.uk/services/research-data/keep-data/lterm-pres

Metadata

provides information on data and the processes of producing and using data. Metadata are data which are needed for proper reproduction and use of the data[20]

Open access

Open access can be defined as the practice of providing online access to scientific information that is free of charge to the reader. In the context of research and development, open access typically focuses on access to 'scientific information' or 'research results'.[21]

Personal data:

any information relating to an identified or identifiable person. A person may be identified by name, images/video, email, IP-address or a number referring to a list of names, through photo/video of recognizable faces, or through a combination of background information.[22]

Research data

may be defined as information relevant to, or of interest to, researchers either as inputs into or outputs from research. They are research materials resulting from primary data collection or generation, or derived from existing sources intended to be analysed in the research project.[23]

Synergies for Europe's Research Infrastructures in the Social Sciences

A Horizon 2020 project focused around three key themes – key challenges facing cross-national data collection, breaking down barriers between research infrastructures, and embracing the future of social sciences. It addresses issues relating to survey design and data collection, data management and curation from a collaborative, cross-national perspective. The project will better equip Europe’s social science data infrastructures to play a major role in addressing the key societal challenges facing Europe today and help ensure that national and European policymaking is built on a solid base of the highest-quality socio-economic evidence. See www.seriss.eu.

[20] OECD 2016: 42

[21] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm

[22] http://www.nsd.uib.no/personvernombud/en/help/vocabulary.html

[23] OECD 2016: 43

References

Connelly, Roxanne, Christopher J. Playford, Vernon Gayle, and Chris Dibben. "The role of administrative data in the big data revolution in social science research." Social Science Research 59 (2016): 1-12.

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

Weller, Katrin, and Katharina E. Kinder-Kurlanda. "A manifesto for data sharing in social media research." In Proceedings of the 8th ACM Conference on Web Science, pp. 166-172. ACM, 2016.