Deliverable Number: 6.9

Deliverable Title: Appraisal/Selection Requirements for New Forms of Data

Work Package: 6 New Forms of Data- Legal, Ethical and Quality Issues

Deliverable type: Report

Dissemination status: Public

Submitted by: CESSDA (NSD)

Hervé L'Hours CESSDA (UKDA)

Darren Bell CESSDA (UKDA)

Sarah Butt ESS ERIC (City/HQ)

Gry Henriksen CESSDA (NSD)

Jindřich Krejčí CESSDA (CSDA)

Marianne Myhren CESSDA (NSD)

Janez Štebe CESSDA (ADP)

Øyvind Straume CESSDA (NSD)

Scott Summers CESSDA (UKDA)

Martin Vávra CESSDA (CSDA)

Date Submitted: August, 2018

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 of Health Ageing 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: L'Hours et al (2018) Appraisal/Selection Requirements for New Forms of Data. Deliverable 6.9 of the SERISS project funded under the *European Union's Horizon 2020 research and innovation programme* GA No: 654221. https://doi.org/10.5281/zenodo.1406926 . Available at: www.seriss.eu/resources/deliverables

Contents

Deliverable Design, Structure and Content

This deliverable 6.9 follows on from D6.7 *Generic high-level workflows for the curation of different forms of 'Big Data'*, by addressing the appraisal and selection requirements for new and novel forms of data as part of Task 6.3 (Connected curation and quality) of the SERISS WP 6: New forms of data: legal, ethical and quality issues.

A note on Common Appendices

Appendices A to D "EOSC, FAIR data, Preservation and the GDPR" are common to SERISS Project Deliverables 6.8 *Versioning requirements for curation and access to new forms of data* and 6.9 *Appraisal/Selection Requirements for New Forms of Data*. Appendices A to C provide an overview of emergent plans for the European Open Science Cloud (EOSC)[1] the wider implementation of FAIR[2] (Findable, Accessible, Interoperable and Reusable) data principles and the European Commission vision of data access and preservation. Several of the references covered have been released since the prior deliverable *Generic high-level workflows for the curation of different forms of 'Big Data'*) was submitted. Together they present the rapidly emerging context within which this SERISS work package and task has worked and which will continue to influence and change the related topics of workflows, versions and appraisal/selection for new and novel forms of data in at least the short and medium term. Appendix D provides an overview of key concepts related to the General Data Protection Regulation (GDPR) which came into force within the timeframe of this task.

Though submitted to the European Commission with each deliverable the Appendices have been assigned a separate Digital Object Identifier (DOI) from the two deliverables.

Introduction

The appraisal and selection process undertaken by archives to evaluate potential deposits against a defined Collections Development policy acts as the gatekeeper for resources accepted for storage, curation, long term preservation and access. The focus here, in keeping with the SERISS participants and audience, is survey infrastructures and disciplinary social science data. The primary focus is on addressing survey data, administrative data, and social media data and, increasingly, linked data products from these sources, but the concepts and criteria applied here are more generally applicable to the evaluation of complex linked data with a range of formats, technical requirements and quality issues.

As identified in the prior deliverable (L'Hours et al., 2018a) we must consider that many factors in 'Big' and otherwise 'new and novel forms' of data are not entirely new to the data archives that are continuously faced with emergent factors related to both traditional file-based research study data and to data not originally conceived with research in mind.

For research data management, archiving and preservation big/new and novel forms of data (NNfD) must be addressed within the context of existing and emerging infrastructures of skills, processes and technologies. The data characteristics themselves, even in the rare cases where they are truly unfamiliar to an archive must be considered alongside new and novel research opportunities and the technologies that support them.

NNfD do present challenges to current curation practice, not least around legal and ethical issues and the V's of Big Data (Volume, Velocity, Variety, Veracity and Value). Change management from traditional to NNfD curation is not only a question of evolving data types. Big isn't always huge but there are elements of NNfD data curation, access and use that cannot be delivered at scale on traditional archival technical infrastructure. New and novel forms of data may not be subject to the same level of formal design and ethical review as 'traditional' study-driven objects and processes intended for research. The process of data linkage from multiple sources derives new 'data products' from existing objects, each with potentially changed risks relating to data quality and disclosure. New and novel sources of

[1] https://eoscpilot.eu/

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

data are associated with new stakeholder relationships, and data provenance and data integrity issues. Though adjusting to bigger, faster, more complex data curation and delivery demands is a standard part of the evolution of data archives' collections one element of this 'big data' revolution is that the technical infrastructure paradigm is evolving to go beyond traditional data stored in databases supported by resource discovery and accessed via web forms. Archives seeking to operationalise 'big data' curation workflows may need to develop or procure access to entirely new technical environments to support an increased demand for linkage and analysis across sources including surveys, administrative and social media data.

The surveys and their survey data are a familiar partners/data type for archives that are undergoing a rapid evolution including increased standardisation (particularly through DDI[3]), compliance with emerging legislation (including GDPR[4]) and demand for use in secondary analysis including increased instances of data linkage. Administrative data suppliers are another known quantity to archives, these data are also increasingly in demand as a component in data linkage but are not consistently developed with research use in mind. Alongside these we have the emergence of social media platforms and increased researcher interest in their data. These platforms are new suppliers/partners for researchers and archives whose data is designed with a commercial rather than a research perspective, is subject to a wide variety of terms and conditions (of access, use and sharing) and whose structure and provenance (including processing history) is more opaque than either survey or administrative data.

These data sources and partnership interact with NNfD in ways which increase the variety of data types (pre and post linkage) to be addressed by archives seeking to identify data for retention and long term preservation. A number of these challenges may change the balance of human-mediated versus machine-mediated archival processes. Where NNfD are 'huge' this presents a driver for archives to scale their systems or seek new partnerships. The velocity of these data may increase the demand for more machine actionable processing. Demands for data may speed up with users seeking near real-time availability or access to continuous data streams. Our trust in the veracity of date depends on provenance information which can be highly variable with NNfD.   All of these factors mean there is a greater degree and a faster rate of variability on data which implies a need for clear and rigorous change management over time. While these data sources are increasingly being managed, there remain few examples of their deposit for long term preservation in archives and their limited availability to date means that it can be hard to identify the future value of the huge quantities of data being generated. Each decision to retain and preserve data has a cost attached which must be weighed against the uncertain losses implied by not choosing to curate and preserve the data for the long term.

While disciplinary repositories are acknowledged as vital centres of excellence to support science they are working in a rapidly evolving data, technology, policy, legal and ethical framework whilst simultaneously addressing the increasing demands of cross-disciplinary science and NNfD.

Though researchers including survey infrastructures and projects are actively working to link survey and administrative population data, and archives are actively engaging with this and with social media data, there are few instances where such data has been formally appraised and selected for deposit with a view to long term preservation. Until a critical number of deposits are reached and sufficient time has passed to validate the archival approaches taken it may be advisable to take a cautious approach to information retention.

Such a delay between the emergence of new data, new technology and deposit within disciplinary repositories is not unusual, but it does place an additional risk on data generated during these periods of change.

[3] http://www.ddialliance.org/Specification/

[4] https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en

Many archives are actively engaging as ‘full lifecycle’ actors to influence their networks of current and future depositors through engagement and training. These efforts embed the need for deposit in research culture and help reduce the ‘downstream’ costs of curation. But until NNfD have developed a level of accepted good practice supported by appropriate curation standards (including standardised metadata, controlled vocabularies/ ontologies and supporting information registries) there will be an inevitable additional pressure on resources at the point of deposit into a Trustworthy Digital Repository (TDR).

This paper provides contextual information relevant to the changes faced by data archives including the rapidly evolving academic and scientific ecosystems and then provides an overview of appraisal and selection practices and possible future approaches to addressing new and novel forms of data.

Contextualising Appraisal and Selection

Evolving Ecosystems

There are a number of emerging plans, proposals and actions relevant to European and global Science Ecosystems that are relevant to planning for new and novel forms of data (NNfD) infrastructure (people, processes and technology) in general, and to the work of SERISS T6.3 in particular (workflows, versions, appraisal & selection).

These developments are strongly integrated with an identified need for audit and certification of data and services (including repository services) in line with the ‘core’ model (CoreTrustSeal, 2016) provided by the CoreTrustSeal as a baseline. Generic high-level workflows for the curation of different forms of ‘Big Data’ (L’Hours et al., 2018a) expands on the notion of a ‘core’ Trustworthy Digital Repository (TDR) approach with higher levels of TDR process rigour and requirements granularity offered by Audit and Certification of Trustworthy Digital Repositories Standard (ISO16363[5]) and Criteria for Trustworthy Digital Archives (DIN31644[6]). Both data and services must be, and be seen to be, trustworthy and FAIR (Findable, Accessible, Interoperable and Reusable).

The developments imply new actors and relationships within the data stakeholder ecosystem and have a direct impact on the organisation context for the SERISS partners as survey and archival actors. Key factors are outlined here while Appendix A provides a more detailed ‘snapshot’ of the stakeholder ecosystem at the time of writing.

The EC Communication "European Cloud Initiative" published on 19 April 2016[7] has three pillars:

- EOSC: the European Open Science Cloud
- EDI: the European Data Infrastructure: a High Performance Computing (HPC), data and network infrastructure
- Widening access & building trust: beyond academic research to include government, industry and small/medium enterprises (SMEs)

The EOSC, HPC and new data and network infrastructures will fundamentally change European science as data infrastructure and services are federalised. The widening of access will broaden the range of stakeholders overall and the range of data users to be served by data archives.

The EOSC declaration envisages an implementation roadmap supporting the move towards:

[5] https://public.ccsds.org/pubs/652x0m1.pdf

[6] http://www.dnb.de/Subsites/nestor/EN/Siegel/siegel.html

[7] https://ec.europa.eu/digital-single-market/en/news/communication-european-cloud-initiative-building-competitive-data-and-knowledge-economy-europe

- Clear governance frameworks
- Definition of the initial services
- Clear business models
- Cost optimisation

Specific actions are envisaged to:

- to develop a better culture of research data management data sharing, and practical skills among EU scientists and innovators, including action on incentives, rewards, skills and curricula related to research data and data science;
- to develop FAIR data tools, specifications, catalogues and standards, and supply-side services to support scientists and innovators, and
- to stimulate the demand for FAIR data through consistent FAIR data mandates and incentives to open data by research funders and institutions across Europe.

The EOSC Model action lines fall into six categories overall:

1. Architecture: reduced fragmentation through interoperability and federation
2. Data: FAIR data management and tools across borders and disciplines
3. Access & Interface: for accessing research data and meeting open data obligations
4. Rules for participation
5. Governance frameworks
6. Services

The defined services are:

- A unique identification and authentication service, and an access point and routing system towards the resources of the EOSC.
- A protected and personalised work environment/space (e.g. logbook, settings, compliance record and pending issues).
- Access to relevant service information (status of the EOSC, list of federated data infrastructures, policy-related information, description of the compliance framework) and to specific guidelines (how to make data FAIR, to certify a repository or service, to procure joint services).
- Services to find, access, re-use and analyse research data generated by others, accessible through appropriate catalogues of datasets and data services (e.g. analytics, fusion, mining, processing).
- Services to make their own data FAIR, to store them and ensure long-term preservation.

In addition to evolving to support NNfD, the archives and other research data lifecycle actors must take account of the risks and opportunities implied by this evolution of wider ecosystems when planning for change.

The Recommendation on access to and preservation of scientific information Brussels, 25.4.2018 C(2018) 2375 final provides one potential approach to highlighting the important of a long-term preservation perspective and offers a mechanism for ongoing transnational monitoring and cooperation. Recital 7 states:

> "Preservation of scientific research results is in the public interest. [ ] Mechanisms, infrastructures and software solutions should be in place to enable long-term preservation of research results in digital form. Sustainable funding for preservation is crucial as curation costs for digitised content are still relatively high. Given the importance of preservation for the future use of research results, the

> establishment or reinforcement of policies in this area should be recommended to Member States."

The federation of existing data infrastructures into the EOSC is envisaged through concrete objectives, progress indicators, implementation plans and financial planning across the key areas of focus:

- Open access to scientific publications
- Management of research data, including open access
- Preservation and re-use of scientific information
- Infrastructures for open science
- Skills and competences
- Incentives and rewards

Structured coordination of Member States at Union level and follow-up to this Recommendation identifies a need for each member state to coordinate the recommended measures and provide a contact point to the Commission towards "better definitions of common principles and standards, implementation measures and new ways of disseminating and sharing research results in the European Research Area". These contact points will ensure *multi-stakeholder dialogue on open science at national, European and international level* with an explicit expectation of systemic, gradual changes in research culture across the relevant actors covering "all research outputs from all phases of the research life cycle (data, publications, software, methods, protocols, etc.).

This work could become a common reference point for aligning national approaches with the support of CESSDA and other ERIC partners.

Legal and Ethical

We may expect ethical guidance such as that provided by funders (e.g. ESRC 2015[8]) to be updated alongside procedures for ethical review boards. Individual researchers must also make personal ethical decision about research, including in response to prevailing public opinion (see below). An individual repository may make local decisions about accepting materials for deposit because they don’t meet ethical criteria, or because prevailing public opinion makes accepting the data a ‘public relations’ risk.

The General Data Protection Regulation (GDPR) is now in force and relevant guidance, such as the Handbook on European data protection law[9] from the European Union Agency for Fundamental Rights is becoming available but there are a number of areas of ongoing change including the position around the Directive on Copyright in the Digital Single Market which is still evolving. Copyright challenges for Archives encompass both the issues of copyright in deposited materials and the need to manage any collaborative platforms which support the upload of content by researchers.

Clarity on some aspects of the GDPR will not be reached until sufficient time has passed for it to be enforced in practice at both the European level and at the level of related national legislation (which has not yet been passed in a number of EC countries). There are also an increasing number of sub-national efforts in the United States to address data protection[10] which will influence non-European data sharing practice. Data stewards will need to identify the level of processing detail which must be communicated to data subjects. Like many other data curators the Archives will work to clarify what constitutes the provision of ‘appropriate

[8] https://esrc.ukri.org/files/funding/guidance-for-applicants/esrc-framework-for-research-ethics-2015/

[9] http://fra.europa.eu/en/publication/2018/handbook-european-data-protection-law

[10] https://www.dataprotectionreport.com/2018/07/u-s-states-pass-data-protection-laws-on-the-heels-of-the-gdpr/

technical and organisational measures' at an appropriate level of transparency.

The requirement for of the CoreTrustSeal on Confidentiality/Ethics states: The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms[11]

In this text we will reference key concepts of relevance to archival digital object management but will not provide recommendations on particular approaches or try to address the remaining factors which will only become clear as organisations seek to implement appropriate technical and organisational measures and systems for data processing (whether consent-based or otherwise).

Consent

Even if there is a tendency for archives to use public task as the justification for ongoing data processing the notion of consent remains critical. The data controller (see Appendix D) remains responsible for ensuring that data subjects are informed of how they are able to exercise their rights and that it is clear which, if any, of those rights are being derogated for the purposes of research or archiving.

Repositories must understand the ethical implications of data types offered for deposit, the prevailing disciplinary standards and, where relevant, capture information to describe ethical evaluations and practices pre-deposit. This may include metadata on ethical review board approval or documentation of appropriate consent, or it may extend to the signed/agreed consent records themselves and any associated guidance (e.g. information sheets) provided to data subjects at the time they consented.

At the SERISS Survey Expert Network[12] workshop Andrew Charlesworth, Professor of Law, Innovation and Society at Bristol University, noted that providing granular assurances to data subjects that data processing would be limited, and then calling on public task as a blanket derogation to justify further, broader use could be seen as actively misleading. This risks reducing the confidence of the individual data subject, of potential future data subjects and of the general public. It is important to differentiate and communicate the difference between ethical consent and consent for GDPR purposes.

Public Opinion

When engaging with personal data, especially when sourced from commercial providers such as social media platforms there may be issues beyond the legal and ethical. The nature of public perception on the sharing of data (Fiesler and Proferes, 2018) may impact a repository decision to accept and make data available even if all legal and ethical assurances have been provided. Recent events, including the Facebook-Cambridge Analytical collaboration[13], mean that both social media platforms and research repositories may have limited appetite to manage negative public attention. This may result in the loss of access to data, or at least the long-term preservation of data with long-term research value.

Technical Infrastructure Context

The available technical infrastructure (whether local to the repository or in a broader partnership context) effects appraisal/selection decisions. These can be broadly grouped into three categories, though any repository may have a mixture of these in place.

Files

Data points are bound into files, often with proprietary structures, ideally in known formats and suitable for long term preservation and re-use by the designated community. File-based curation has clear implications for versioning (L'Hours et al, 2018b). Access is often possible via direct download by users. File based data does not preclude structural complexity or the

[11] https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf

[12] https://seriss.eu/resources/sen/

[13] https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal

need for clear objects/grouping models to handle relationships (e.g. between data sets or from data to related publications).

The traditional notion that all materials relevant for re-use are curated and made available by the archive may not always apply with NNfD. In the case of a Twitter dataset the terms and conditions of use may limit the archive to maintaining only tweets´ identifiers. These 'dehydrated' files may not be sufficient on their own for ongoing research. Such 'distributed objects' (see below) where elements are not under the control of the repository is a feature of NNfD and co-dependent data infrastructures. This may complicate re-use and affect the reproducibility of research

Databases

Curation and delivery of more granular data, addressable to the variable or question level, is often undertaken through SQL or other traditional database/structured data paradigms, such as XML. Delivery may be through web-accessible interfaces which offer a range of query, analytic and visualisation options. The systems can offer support for linkage of data from multiple sources.

Native Linked Datastores

Structured data may also be curated through resource description framework serialisation (RDFa,[14] JSON-LD[15] etc.) to make it amenable to linkage with a wider range of similarly structured data, across multiple infrastructures. This supports exploratory data analysis at speeds and scales not practical with file or database driven data environments.

A repository may receive a dataset in SPSS format, enrich and curate it through a SQL database for use in online analytics environments in addition to using datastores. RDF data systems may be needed to support evaluation and quality control of large complex datasets, as part of the appraisal and selection process, as data access points, or to support 'round-tripping' of data into multiple access and preservation formats.

Object Models

A common vision of digital objects and object characteristics across depositors, repositories and data users is critical to consistent management of traditional and NNfD. Object models support clear versioning and management of related instances of data assets such as anonymised versus sensitive versions of datasets (L'Hours et al, 2018b).

Rights Models

A move towards national and international interoperability of infrastructures has created some alignment of licences and access management, but conditions of deposit and use often begin life as prose, which is reviewed by a legal department and then instantiated as textual licences and access management source-code. A unified machine-actionable rights model (incorporating IPR) will be required to enable the security and portability of NNfD. One candidate for such a model is the Open Digital Rights Language (ODRL), which provides "a standard description model and format to express permission, prohibition, and obligation statements to be associated to content in general"[16]. A variety of efforts to support data privacy are underway including those to address data tags and differential privacy[17].

Training & Knowledge Base

As NNfD data are deposited in archives supporting multi-domain/multi-disciplinary data the underlying data and technology knowledgebase of repository staff must expand.

Both the EOSC development and the focus on FAIR data refer to the need to support and

[14] https://rdfa.info/

[15] https://json-ld.org/

[16] https://www.w3.org/TR/odrl-model/

[17] https://privacytools.seas.harvard.edu/

train researchers but this seems to remain disconnected from public information to support data literacy and communicate the potential benefits of sharing personal data for research.

Actors and Complex Partnerships

Within the wider group of research data lifecycle stakeholders (including legislators and standards bodies which effect the ecosystem) a subset of parties are more directly involved in our workflows. These 'actors' may be individuals, organisations or machines (sometimes referred to as 'agents') and the number of parties are likely to expand as we seek to support a wider range of data sources with more varied technologies delivered to a wider user-base. Wherever possible the actors responsible for research-related data handling should be defined in advance (OECD, 2016).

At the point of appraisal and selection, decisions define what curation and access workflows are required from which actors. The OAIS Reference Mode for an Open Archival Information System (OAIS, 2012) framework identifies two key actors: data Producers (Ingest) and data Consumers (Access). The term 'Producer' may also be a proxy for the actor undertaking deposit, whether on behalf of a publicly financed research project, a commercial entity or a government department. The data producer and repository together identify the appropriate curation activities, access conditions and any embargo periods which are instantiated in licences. There is also a need for awareness of 'pre-repository' actors that have influenced the data and metadata being deposited, and of the expected application of data at the point of access/use.

There are various perspectives on these actors (Lyon, 2007) e.g.[18]

- Scientist: creation and use of data
- Research institution: intermediate short-term curation of and access to data
- Data centre: long-term curation of and access to data
- (re)User: use of 3rd party data
- Funder: set/react to public policy drivers
- Publisher: maintain integrity of the scientific record
- Science policy maker

Actors may fulfil multiple roles, for examples ERICs may be scientist, data centre, and publisher and are involved in both data collection and research governance.

Actors' influence on the data/lifecycle can range from an effect on the data quality, accessibility and use, to those that influence 'research governance' (including legal, ethical and privacy issues), assignment of data controller/data processor responsibility, and those representatives of the scientific community involved in peer review and ethical review.

Archives, Repositories and other Infrastructure Actors/Entities

Repositories are data stewards who undertake responsibility for the curation of data. A trusted digital repository (TDR) additionally meets certain standards and assures (whether directly or indirectly) that the data under curation is preserved (retained usable and understandable to a designated community-described further below) for the long term (in effect beyond the next round of changes to the data supply or demand environment).

Beyond this designated community data may be accessed by an increasing range of users with less knowledge and experience of the data, methodologies and technologies. The repository must decide if/how to support policy makers, journalists and 'citizen scientists' at the point of use. There may be considerable costs to contextualising or presenting data to this wider user-base.

A TDR is primarily a disciplinary/cross-disciplinary collection of data experts working in the

[18] Liz Lyon: Dealing with Data: Roles, Rights, Responsibilities and Relationships - Consultancy Report. http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.doc

post-data creation/collection phase. The current and future business processes can be controlled through a process of clear workflows management (Georgakopoulos, Hornick, & Sheth, 1995) involving:

1. defining workflows, i.e., describing those aspects of a process that are relevant to controlling and coordinating the execution of its tasks (and possibly the skills of individuals or information systems required to perform each task), and
2. providing for fast (re)design and (re)implementation of the processes as business needs and information systems change

Survey Projects and ERICs

Survey ERICs and projects are data infrastructures in their own right. They must align with the legal requirements of the GDPR, including the transparency requirements and the provision of appropriate technical and organisational measures, and will be subject to the rules of participation in the EOSC which call for trusted repositories and FAIR-certified data.

Under the outsource options offered by the CoreTrustSeal actors in the data collection and analysis phase of the lifecycle can seek to become TDRs if they can demonstrate that they have a partner to take long-term preservation responsibility for their data.

Designated Communities, Service Users and Stakeholders

The OAIS, ISO16363, DIN31644 and CoreTrustSeal are all based on fundamental assumptions about disciplinary/domain repositories delivering long term preservation services (Giaretta et al., 2012) around data in which they are experts. Central to this service delivery is the notion of a Designated Community:

> An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the Archive and this definition may change over time (OAIS, 2012).

The concept of 'designated community' guides decisions on how the data needs to be presented and accessed, based on the 'knowledge base' necessary for users to independently understand the data. Alongside the designated community is a larger potential audience for the data services offered by the repository including policy makers, educators, journalists, citizens' scientists and the public at large, whose demands may help set criteria for which data to acquire and make available. For instance a recent strategic review of longitudinal data (Davis-Kean et al., 2017) recommended:

> "Development of a centralised analysis platform aimed at policy users of its longitudinal data resources to access information related to these resources and data enhancements, and facilitate analysis. This dashboard would provide descriptive statistics and share data in a way that is accessible to users who are interested in longitudinal data, but with diverse interests and varying levels of statistical knowledge and methodological training."

Delivery of data and services to a wider audience will depend on the priorities of the repository funding bodies but it is important to maintain the distinction between these users and the designated community. It may be useful to build these considerations into stakeholder identification and management, for example:

Designated Community: Social Scientists for whom the data must be usable and understandable

Targeted User Communities: funders, policy makers, journalists and others who benefit from data and other resources and to whom the archive must demonstrate impact

Users: the superset of individuals who access the archive's site, data and services

Other Stakeholders: potential future members of the target, user, or designated communities

Curation for full understandability by non-disciplinary experts can be complex, time consuming and expensive. As cross-disciplinary research becomes more common a certified repository must decide whether to amend their designated community and will need to provide supporting evidence that, for example, a social science data archive has the expertise to support related medical data.

Outsourcing and Complex Partnership Systems

Tracing data and analytic provenance across complex 'big data' platforms and knowledge production ecosystems, while addressing data privacy issues is a challenge. But in addition to complex routing and dependencies across the digital object lifecycle what is perceived externally as a single archive, repository or data centre is increasingly likely to be a complex partnership of actors and services.

This presents challenges to certification efforts trying to audit and approve a clearly bounded entity. Repositories curating traditional file-based resources already have access to a range of products and services to support their workflows. The transition to more complex hosting, storage and access partnerships and reliance on registries (formats, metadata) managed by others presents a challenge to ensuring the integrity, security and provenance of data and implies the need for clearly documented chains of custody even within the repository phase.

Appraisal & Selection of New and Novel Forms of Data

The text above provides context for the challenges and opportunities facing repositories and other data infrastructures in a fast-changing legal, technical and partnership environment. This and following sections provides an overview of appraisal and selection and some of the key issues that guide the process including archiving goals, meeting the needs of user communities, NNfD data sources/types, costs and data management plans.

The Appraisal and Selection process[19] against a defined Collections Development policy acts as the gatekeeper for digital assets (data, metadata and documentation) accepted into data archives (CESSDA service provider or other similar data service infrastructure) for storage, curation, long term preservation and access. Appraisal and Selection falls into the Pre-Ingest function of the archival phase and must consider which characteristics of data, including supporting metadata, technologies and skills, affect the decision to accept data for deposit. Assessment includes potential for re-use and the availability of predefined workflows and services to support the specific data types. Based on the outcome of those assessments decisions about data management options are taken.

One approach to developing new paradigms of appraisal and selection for these data types is to reassess the traditional appraisal/selection approach with new and novel in mind. The existing paradigm is based on evaluation of data as to their value through alignment with collections development criteria. This can be extended to cover the properties of NNfD, or changed where there is a qualitative shift.

With NNfD the set of actors with whom to communicate broadens: producers of data, and through them the data subjects themselves, data owners, secondary users (with added value products, or 'statistical analysis ready data' related to publications), all have a role. A dialogue can be established between actors in the process, which may lead to bilateral agreement, followed by data deposit or harvesting based on agreed criteria. E.g., a repository may reach an agreement with a social network company to store and make available the data for public (including research) use, but it must be acknowledged that at present the archives have few established relationships with these potential NNfD suppliers

[19] CPA2.1 Data Acquisition and Ingest; https://www.cessda.eu/Projects/All-projects/CESSDA-SaW/WP3/CESSDA-CDM/Part-2-CRA2-Digital-Object-Management/CPA2.1-Data-Acquisition-and-Ingest

and limited influence over the data production process.

The starting point in designing an appropriate paradigm is to develop a set of dimensions and corresponding appraisal/selection criteria pertaining to a broader set of types of data. One of the dimensions would be legal/ethical characteristics of data, which is less clear in data products where 'passive participants'[20] contribute. If personal data is reported or detected at the negotiation/deposit stage restricted licence access conditions may apply or, the data may be processed to remove the restriction, for example through anonymisation. A variety of access routes may to be offered, including some where the repository continues to mediate use and holds approval right before data is released, such as statistical disclosure control. These access modes include safe room, secure remote access, download, variable browsing etc. The service provider needs to evaluate the cost of running any additional services, including training costs that would be needed for certain of those modes (e.g. familiarity with the rules of conduct for secure access data). If specific workflows and services cannot be established at the repository, this then precludes the selection of certain data types in the repository. Over time as demands emerge for the deposit and use of NNfD repositories need to consider developing new supporting services, whether individually or collectively e.g. some may specialise in offering secure access services to the larger community.

The decision process involves a variety of data lifecycle actors/perspectives including: data producer/rights-holder, end-user, and the data archive itself. Long-term research potential is one basic criteria for selection and appraisal; the outcome that we seek is to ensure data of value to our designated community remains accessible and usable. The costs associated and the available expertise of data stewardships needs to be reflected in this workflow as we seek to find the optimal strategy in dealing with data("Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access," 2010) (Blue Ribbon Task Force, 2010.

> 'Domain repositories in the social and natural sciences each serve a scientific community, which may be a traditional academic discipline, a subdiscipline, or an interdisciplinary network of scientists, united by a common focus. This in-depth knowledge enables domain repositories to enhance the data ecosystem far beyond data preservation and access. By combining domain-specific scientific knowledge, expertise in data stewardship, and close relationships with scientific communities, domain repositories accelerate intellectual discovery by facilitating reuse and reproducibility, ultimately building an enduring record that represents the richness, diversity, and complexity of the scientific enterprise.' (Ember & Hanisch, 2013)

The appraisal and section archival function has an analogous process in the research phase of the lifecycle. In a research setting decisions about finding and selecting of existing data, or the production of new data, is based on a cost/benefit analysis of its' ability to address an immediate research purpose. Thus, for individual research projects the question is to find or produce the data that addresses a specific research question, including consideration of data quality and acquisition cost.

For archives there are similar considerations, but the evaluation must integrate the potential for generalised reuse of data in different scenarios by different end-users. Costs associated

[20] "Passive participants" typically do not directly engage with the researcher and may not be aware of the current research. This includes those whose behavior, social media use, or other public activities are observed, those for whom information is obtained through analysis of publicly available information, or those whose existing research or administrative records are used under previously established provisions that allow for their use in the current research without need for additional consent. (http://www.aapor.org/Standards-Ethics/AAPOR-Code-of-Ethics.aspx)

with deposit, curation, preservation and access are balanced against the perceived utility of data: 'Appraisal is a key component of Long Tail data management – data preservation and dissemination has a cost and not all data will be preserved.'(Genova & Horstmann, 2016)[21] Yet, for the long-term, '(s)election decisions should reflect the interests of future as well as current stakeholders; (…) it is vital that the proxy organisations have well - articulated and transparent selection procedures, and that they make full use of domain-specific expertise to advise on the selection and preservation of materials.' (Blue Ribbon Task Force, 2010)[22]. With large volumes and a large variety of data, the key consideration is 'what is the combined cost of creating and managing metadata, and undertaking preservation?'[23]

High level workflows (L'Hours et al) presented the standards and audit model which underpins "OAIS" archives, subject to Trustworthy Digital Repository (TDR) certification, serving data preserved for the long term to designated communities. It was identified that for these bodies, adapting to new user demand for new data from new sources, requiring new technologies is fairly standard operating procedure. This is not to claim that these reference models and audit standards are complete, or that all repositories have attained the ideal level of maturity, but only to demonstrate that most challenges implied by NNfD are familiar to longer-standing archives. There remain considerable challenges at this stage of considerable cumulative change. The archival response needs to integrate new ecosystems, technologies and research opportunities as well as the data themselves.

To best engage with new opportunities and mitigate risk, the appraisal and selection process must take a 360 degree view, covering not only the local organisational context but also the wider ecosystem and the pre- and post-repository stages of the (research) data lifecycle. Where has the data been, what is it for, where is it going?

In all data archiving, but particularly for data sources which have yet to demonstrate long term value initial appraisal is not sufficient. Ongoing periodic re-appraisal is required to justify ongoing retention. Data stored under the precautionary principle may not have the longer term value expected, data accepted for a lower quality of curation (e.g. bit-level integrity assurance) may be identified as of greater value than initially thought and would benefit from higher levels of long-term preservation stewardship. Over time alternate locations or suppliers may emerge, able to offer improved services around the data. In a connected ecosystem there is a lower risk of destruction of data (or of insufficient stewardship increasing the long term risk to access) because it is easier to identify, interact with and manage custody transfers with alternate data curators. Appraisal and selection also necessitates an awareness of the options available in the wider ecosystem and processes to point data rejected by one repository towards another environment.

Working from the context section above we can identify a number of features for consideration at the point of appraisal which will impact not only the actions of the receiving repository, but also the wider ecosystem.

[21] New and Novel forms of Data (NNfD) in the social science setting share many of characteristics of so called Long Tail research data, and can be treated similarly if not taken as synonymous.

[22] Pp. 76, Proxy organisations are considered those that provide services to support long term public interests, including future research use, against the organisations that support short term goals of particular project or organisation itself. In the context of social media data, commercial social media companies are ill-suited for the valuation and preservation of content that is generated on the platforms.

[23] http://www.dcc.ac.uk/resources/how-guides/appraise-select-data. Compare Instalment on "Appraisal and Selection" http://www.dcc.ac.uk/resource/curation-manual/chapters/appraisal-and-selection/ Ross Harvey and http://www.unesco.org/new/en/communication-and-information/access-to-knowledge/preservation-of-documentary-heritage/digital-heritage/concept-of-digital-heritage/

Archiving Goals

> Different kinds of digital materials, created in different contexts for different stakeholders, require different approaches to appraisal and selection. For instance, the conditions under which data are acquired by a national library as part of legal deposit legislation are different from scientific datasets, and they both have different characteristics from records of business transactions created in digital form, and so on.[24]

A legal deposit requirement may require that a national archive or library receive all published outputs or governmental records. This type of mandated deposit may not be a feature for disciplinary repositories but for example, research funded by the UK ESRC has a mandate to offer data for deposit while the repository retains the right to apply a selection and appraisal process. In other cases, researchers seek out the repository, perhaps as recognised expert in their disciplinary field.[25] Repositories also actively seek data which aligns with their collections policy.

The intended purpose(s) for which the data will be used effect appraisal decisions and influence repository workflows. Data archiving to support data citations in publications is increasingly common, and may be mandated by the scientific publisher (Cousijn et al., 2017)[26] but it is not always clear what level of curation the cited data undergoes. Detailed metadata on population sampling, methodologies weighting etc. may be sufficient to repeat a study, but for reproducibility the original data must be available and remain usable.

Some outputs, such as those provided by some major survey may have been developed as reference datasets with ongoing value and re-use built into their development. In other cases, project outputs may be archived as 'good practice' but their re-use value may be limited, or may only be realised through extensive curation and enrichment.

Data Sources, Types and Characteristics

Personal Data

A primary consideration during appraisal and selection is whether the deposit includes personal data, including pseudonymised data or anonymised data which may be at risk of re-identification during re-use. If so the repository must immediately consider appropriate storage and protection of the data, or of sample data provided for evaluation. The presence of personal data will dictate a number of key curation workflows. Even for data which are considered to have been fully de-identified and anonymised the involvement of data subjects will imply a need to:

- Receive guarantees of anonymisation
- Undertake additional repository checks to ensure this (if resources permit)
- Gather appropriate evidence of ethical approval
- Identify that consent has been given and assess any limitations on use/re-use implied by that consent.
    - And/or, identify exemptions from the rules of processing personal data (e.g. publicly disclosed data)

Survey Data

Survey data from survey ERICS or projects are a traditional form of data for data archives with well-developed workflows available. Survey projects often have long-standing deposit relationships with archives and are subject to ethical and methodological review throughout their funding applications, development and implementation. Data are often generated with the expectation of general research re-use as well as to address specific research questions. The level of general applicability of the data will influence the curation necessary to make the

[24] DCC Curation Reference Manual. http://www.dcc.ac.uk/resource/curation-manual/chapters/appraisal-and-selection/ , pp. 12.

[25] https://www.nature.com/sdata/policies/repositories

[26] https://www.biorxiv.org/content/biorxiv/early/2018/03/07/100784.full.pdf

data understandable by re-users. In the case of longitudinal work there is an inherent assumption of longer term use which supports the preservation mission of archives. Survey actors are strongly engaged with the research community and their needs.

Administrative Data

> Administrative data are defined as data which derive from the operation of administrative systems, typically by public sector agencies. They cover activities such as health maintenance, tax and social security, housing, elderly care, vehicle and other licensing systems, educational progress, etc. While such data are not designed for research purposes, they often have significant research value, especially when linked to other datasets or to user-generated surveys. (Duşa, Nelle, Stock, & Wagner, 2014)[27]

Administrative data, with population data being a common example, is not typically developed with linkage and re-use by third parties in mind. But these data may have analytical potential for evaluation of public policies[28]. Design, development and change over time may not be documented to the same level as for survey data. Perceived risks around third-party hosting and use of data, including public opinion and the risk of disclosure or identification may be a barrier to administrative data holders choosing to share these data.

Social Media data

As noted in the SERISS Versions document (L'Hours et al, 2018b) social media data is created to deliver commercial services which may be added, changed or removed rapidly and frequently in response to user demand and to public perception. Data is not created with research in mind. Underlying data schemas are not transparent and legal data access modes (where they exist) may not support consistent repeatable querying or selection of data as representative of a particular population (Kinder-Kurlanda et al, 2017). Terms and condition may also change, presenting a challenge to maintaining compliance when data has been deposited in an archive. In the case of Twitter the limitation of depositing 'dehydrated' tweets impacts the integrity of the dataset and rehydration at the point of access may not be complete (e.g. absence of deleted tweets). In effect the repository is holding part of a 'distributed object' which is added to existing issues with a lack of available information as to how Twitter data is collected, stored, cleaned and analysed (Driscoll & Walker, 2014).

Data Linkage modes

Data Linkage undertaken prior to deposit has less of an impact on repository decisions as the derived data products can be addressed in the same way as most data offers.

Additional documentation and metadata is required when, for instance, linked data have different levels of provenance or perceived technical data quality (e.g. survey data vs. social media data).

The main impact of linkage on appraisal and selection is if deposited personal data will be linked to other data in an environment controlled by the repository. In this case there must be consideration of the potential for re-identification. Though there are acknowledged benefits from data linkage the upwards trend in this approach may be limited by data security and

[27] Peter Elias: Administrative Data. In **Facing the Future : European Research Infrastructures for the Humanities and Social Sciences.** http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:kobv:b4-opus-26346**;** See also definition at Administrative Data Liaison Service (http://www.adls.ac.uk/adls-resources/guidance/introduction/), superseded by *The Administrative Data Research* Network *(*https://www.adrn.ac.uk/) where ADRN GLOSSARY contains 'Definitions of Terms used in the Network' (https://adrn.ac.uk/media/173982/adrn018-glossary_00_09_pub.pdf)

[28] Institute of Labor Economics (IZA). IZA/IAB Linked Evaluation Dataset (yyyy – yyyy years of data suing in your analysis). International Data Service Center of IZA (IDSC). Version 1.0. doi:10.15185/izadp.8337.1

privacy concerns (Künn, 2015).

Interdisciplinary data

Technologies which support more advanced data linkage and the opportunities presented by data sourced from multiple sources have driven the demand for access to cross-disciplinary data as well as to administrative and social media. For domain/disciplinary repositories the local knowledge base (or access to appropriate external expertise) must keep pace with moves into cross-disciplinary work. Factors include an understanding of data collection practices, methodologies, terminology and unfamiliar formats, all of which impact the ability to evaluate, curate, contextualise and provide access to data with a cross-disciplinary source.

Broadening an archives' scope into new disciplines also apply new data formats to be understood and evaluated as acceptable for deposit, preservation and access.

Preservation

Preservation of a clearly bounded file or collections of files alone is becoming less common. Even basic metadata can depend on schemas developed and maintained by third parties and the opportunities of linked data are parallel by a range of challenges as to how the 'related' objects can be preserved. These include dependent registries and standards but also software/code used in the development of the data. Whether long term preservation is assured by format migration or emulation of original environments not all of these related/dependant artefacts will be under the control of the archive.

A broad view of costs

Limitations on resources mean that not all of the vast quantities of data generated by people, governments, non-governmental organisations and commercial actors can be actively curated for the long term by domain/disciplinary experts. A cost-benefit analysis, dependent on assigning a value to data, is implied even if not literally undertaken.

The costs related to the selection of any dataset for deposit must be weighed against the collection development policy, mission and the need to ensure sustainability of repository services. There are several forms of associated cost. There is a societal cost: the exposure of the population and individual data subjects in the ethical and privacy protection framework, balanced against the costs of additional burden if the same data needs to be collected or generated anew.[29] Criteria similar to those that support decisions about data collection for narrow research purposes can also apply in selection and appraisal: the relevance of data for advancing knowledge, applicability to addressing societal problems etc. The potential future applicability of NNfD data for research purpose can be harder to assess, as these 'organic' data are not primarily intended for research. Experience has not yet been accumulated so archives lack examples of long-term value.[30]

Repositories are aware that their general-purpose disciplinary knowledge of the data and required curation best practice is balanced against the specific knowledge of the researchers involved in study design, data collection and analysis. Over time metadata collected at deposit has been expanded to capture as much of this pre-repository knowledge as possible. Repositories have also been active in designing and delivering training in research data management (RDM), often based on archival experience, to social-scientists. RDM improves practice upstream and supports pre-repository capture of relevant provenance knowledge. RDM also reduces capture and curation costs downstream, for repositories.

Though it is (currently) harder to identify NNfD value in the long term, any decision to retain them (even without preservation actions) has cost implications. NNfD also imply increased

[29] If at all possible, considering that one of the characteristics of Big data is that it usualy contain the whole population of units of observation that is unique in time. If this is lost, it can't be reproduced again. This, however, dose not mean that the whole target population is covered.

[30] https://www.census.gov/newsroom/blogs/director/2011/05/designed-data-and-organic-data.html

costs to trusted digital repositories as the lack of standardised RDM practice and provenance will increase deposit and curation costs. Over time new curation and RDM practice will evolve and costs will reduce and move back upstream. In the meantime, appraisal and selection processes may consider maintaining more NNfD as a precautionary measure only if resources are available.

Data Management Plans

A key feature of academically derived data in contrast to administrative and social media data is the increased likelihood that it will be deposited with a data management plan (DMP). Principles (Michener, 2015) for creating plans are supported by online tools[31] and there are active efforts[32] to make these living, versioned and machine actionable.[33]There are a number of sources of guidance for data management planning including the CESSDA Expert Tour Guide on Data Management for social science researchers.[34]

The expansion of DMPs to consider NNfD and the implications of personal data and data linkage would be a great advantage. Early identification of data characteristics (personal data, specialist quality control, cross-disciplinary) may help dictate future lifecycle routing and whether the infrastructure and services of a repository align with the DMP.

For NNfD whose long-term preservation is not yet standardised and whose long term value is not yet realised it may be valuable to design a DMP where a TDR undertakes speculative acquisition of the data i.e. stores and curates the data as though it will be preserved for the long term, but with the intention of periodically re-appraising the data for long term value.

Demands from within the scientific community for research transparency, trusted repositories and FAIR data in collaboration with data archives, scientific journals and funders, mean that more entities and services will be integrated into the research data lifecycle. As many of these actors, roles and responsibilities should be captured in the DMP as possible.

Collection Development Policy

A documented approach to collections development or acquisition guides the appraisal and selection procedures in the data centre.

> An established set of criteria will aid the repository in determining the types of information that it is willing, or required to, accept. This is necessary in order to make it clear to funders, depositors, and users what responsibilities the repository is taking on and what aspects are excluded. It is also a necessary step in defining the information which is needed from the information producers or depositors […]. The ingest process sees quality checks performed on the data and documentation and adds information (value) that is relevant for the preservation, discovery, and reuse of the data. [35]

The collection policy also needs to specify which external actors it is responsive to. For many social science data archives in a mature data sharing culture, this includes deposit mandates from the primary funder with 'Formal agreements in place, either with government,

[31] https://dmponline.dcc.ac.uk/

[32] https://www.rd-alliance.org/groups/dmp-common-standards-wg

[33] https://blog.dmptool.org/2018/07/09/scoping-machine-actionable-dmps/

[34] https://www.cessda.eu/Research-Infrastructure/Training/Expert-Tour-Guide-on-Data-Management

[35] CPA – Capability process area; RA – Required actitivities; See https://www.cessda.eu/Projects/All-projects/CESSDA-SaW/WP3/CESSDA-CDM/Part-2-CRA2-Digital-Object-Management/CPA2.1-Data-Acquisition-and-Ingest

funders or research institutions (e.g. funded researchers are obliged to deposit data at the repository)".[36] Funders may expect the data used and/or produced in publicly funded research projects to be deposited at the disciplinary data centre. It may specify the requirements that grant holders need to follow, such as that the data from publicly funded project needs to be deposited in the recommended repository.

Such requirements define the types and sources of data the repository needs to accept by default e. g. any data that is used or produced during publicly funded research project needs to be considered for acquisition, including the NNfD, which potentially challenges the established acquisition criteria of the disciplinary data centre. The data centre would need to assess whether the data fits the primary collection policy, including quality criteria. If not, a self-archiving service could be offered as a substitute, or advice about where else the data can be deposited to fulfil the funder's expectations. Each of the decisions needs to be covered by established criteria and procedures.

The challenge of NNfD is that either new services need to be established that cover the extended variety of data types, or a service is established with a broader scope and more loosely defined requirements for data, documentation and metadata content and format.

Most acquisition workflows are guided and improved by monitoring designated community needs. These are ideally then translated into the collection policy statements, and from this procedures and workflows are derived. The policy typically consists of specifications about which types of data the organisation accepts, processing levels, and points to specific workflows that specific types of data would follow.

The sub-headings below, derived from the CESSDA CDM[37] Data Acquisition and Ingest section reflect a number of artefacts, processes and services that may need to be reviewed with NNfD in mind.

Documentation/ Metadata requirements

What are the data types and characteristics? Has data been linked, or is it suitable for linkage? Does it have TEI[38] or some other metadata formatting for Social media data? Does it support designated community knowledge base? Is the data multidisciplinary? Is recommended metadata (e.g. specified by CESSDA) sufficient? Does it require an update to processing or delivery technologies?

Collection policy

Does the data align with the current collection policy? If not, does the policy need to be updated to address the NNfD?

Systems for submission

Are deposit devices, interfaces and procedures adjusted and optimised based on technology watch, monitoring of, and communication with the Designated Community and other relevant stakeholders? Does the offered data reach any size or throughput limits? What other technical requirements are there for different types of data which needs to be fulfilled? Technology requirements, versioning, continuous flows of data, automation of procedures? Security? Format? Large collection of digital objects in different formats may be accepted as already annotated in a proprietary data analysis tool. To extend the qualidata analogy, the Qualitative Data Repository (QDR), Center for Qualitative and Multi Method Inquiry (CQMI) accepts Computer-Assisted Qualitative Data Analysis Software (CAQDAS) formatted data that stores a wide range of objects and references either in proprietary format, or exported in a raw format.

Authentication and authorisation

[36] . CESSDA Maturity Model: activity *RA2.1.1.1: Methods for acquisition and selection of data* to achieve the 'Defined level',

[37] https://www.cessda.eu/Projects/All-projects/CESSDA-SaW/WP3/CESSDA-CDM

[38] http://www.tei-c.org/

How are the data producers, depositors and rights holders identified? Are they: registered researchers, primary data owners, commercial, public sector or members of the general public?

Requests for provenance information

Are provenance requirements set corresponding to the NNfD types? Have challenges in documenting search and sampling procedures been incorporated?

Citations

Is there an unique citation available, related to a global PID policy/system for the specific features of the NNfD (may be accessed from different locations, versions and updates are common, current data flows, secured access data with limited provenance trace, data that was deleted after linking, etc.)

Conditions placed on content, deposit licenses

Does the set of conditions need to expand for new data types?

Legal transfer of custody; agreements on rights/responsibilities

The conditions of data deposit depends on agreement with the data producer on terms of service (ToS) and agreement about IPR e.g. licence or contract. Data that contains information about individuals must be supported by consent metadata and stored and accessed under appropriate conditions.

One strategy to overcome the obstacles in the third party data would be to have an agreement with the original data producer (or owner). Again, following the analogy with the qualitative data in social sciences, such agreements are sometimes made on the national level regarding the research use and archiving of licenced content as enacted in agreements in Finland[39].

Receipt or request to resubmit

Is the process partly or fully automated? Are quality assurance criteria well established, consensus reached that would allow to communicate back if update is requested?

Quality Assurance

How is completeness and correctness defined, along with other quality assurance criteria? With a wider variety of data sources, some of which are only implicitly structured, it can be difficult to develop thresholds for formal quality requirements.

Authenticity checks of deposited material

The full history of the data may be absent. Some parts of the digital object may be stored elsewhere (cf: Twitter dehydration) beyond the control of the repository. Some administrative data may be retained within the creation environment. Can the archive identify is data has been deleted or amended in the pre-repository phase?

Quality control standards and reporting mechanisms

Manual or automatic? Formal features only or scientific value assessment (validity, reliability, error)? Alongside the questions above relevant to redeveloping a Collections Development Policy for NNFD there is a key requirement from the CoreTrustSeal[40].

“Appraisal: R8. The repository accepts data and metadata based on defined criteria to

[39] ‘The FSD can also archive newspaper and magazine material as well as photographs, cartoons and illustrations in books that have been collected by researchers for their studies but created by someone else. According to an agreement between the FSD and the Finnish copyright society Kopiosto, the data archive can archive and disseminate such material for research purposes.’ http://www.fsd.uta.fi/aineistonhallinta/en/processing-qualitative-data-files.html#data-collected-from-periodicals

[40] https://www.coretrustseal.org/

ensure relevance and understandability for data users.

The appraisal function is critical in determining whether data meet all criteria for inclusion in the collection and in establishing appropriate management for their preservation. Care must be taken to ensure that the data are relevant and understandable to the Designated Community served by the repository.

For this Requirement, responses should include evidence related to the following questions:

- Does the repository use a collection development policy to guide the selection of data for archiving?
- Does the repository have quality control checks to ensure the completeness and understandability of data deposited? If so, please provide references to quality control standards and reporting mechanisms accepted by the relevant community of practice, and include details of how any issues are resolved (e.g., are the data returned to the data provider for rectification, fixed by the repository, noted by quality flags in the data file, and/or included in the accompanying metadata?)
- Does the repository have procedures in place to determine that the metadata required to interpret and use the data are provided?
- What is the repository's approach if the metadata provided are insufficient for long-term preservation?
- Does the repository publish a list of preferred formats?
- Are quality control checks in place to ensure that data producers adhere to the preferred formats?
- What is the approach towards data that are deposited in non-preferred formats?"[41]

Until best practices in the appraisal and selection of NNfD are established it will be challenging for repositories to meet a number of these criteria. A key point when considering the collection policy for NNfD is monitoring of the demand and the supply of preservation services in the future (Blue Ribbon Task Force, 2010)[42]. In particular, *'quality control standards and reporting mechanisms accepted by the relevant community of practice'* are not yet established. It will take time for all research actors to gain sufficient experience with NNfD to inform best practice in archiving.

Appraisal/selection criteria

Appraisal and selection criteria are used by an expert committee[43] to review offered data in relation to:

- the content of the study,
- methodology,
- scientific relevance,
- legal consistency
- documentation

The resulting decisions need to be sustainable and cost effective. This often means that different processing levels are assigned to the 'data collection'[44]. This can contain different

[41] https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf

[42] 'Whatever preservation strategies are used, demand for preserved information must be articulated well enough to ensure there is sufficient supply.' '(…)understanding remains provisional in matters concerning new genres, such as social networking sites, and in managing scientific data produced on a scale that is unprecedented.'; pp. 18, http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

[43] E.g. https://www.ukdataservice.ac.uk/deposit-data

[44] UK DS Collections Development Policy, 19 January 2016, Version: 05.00, available at:

preservation strategies (e.g. bit level storage versus active preservation through format migration), and different time scales.

Selection and Appraisal Criteria usually fall under the general set of categories, which are limited in scope reflecting the ability of the repository to offer the customised data management service. Criteria used to review data and documentation assess the value of the data, in relation to key actors' needs. Some of the following general criteria that are used in various data services may have a specific meaning concerning NNfD

Relevance to Mission: Based on funder requirements, organisational strategy and the levels of production and use of NNfD in the disciplinary research setting.

Relative Scientific or historical value: Some NNfD data are historically unique, e.g. social media data collected around national elections or referendums related to important populations, such as members of parliament. The decision to accept such data is weighed against the possibility that other institutions are already holding the data and/or can provide a more sustainable service. In evaluating the digital heritage value of social network data (Twitter, Facebook, web…) the question of which actors are taking responsibility may be best discussed at a peer group level e.g. within digital humanities and digital heritage working groups. For data that was generated for a specific purpose the repository must consider whether the research potential has been exhausted by the original researcher.

Uniqueness criteria: a resource is not available elsewhere, and not readily reproducible. New sources or types of data are specifically mentioned as a type of data that enrich the data archive collection by offering researchers a variety of sources from which to choose. The data may be interesting if it is linked or contextually related to some other data, e.g. ESS media reports are related to the main survey collection.

Educational value, value for the discipline: Studies that are interesting for learning new methods, or historical studies that are important in a given field. This is usually one of the criteria for Qualitative data selection, when the wider usability of data may be disputed: "the relative importance or impact of the study e.g. research recognized to have had a major influence in its field and/or representing the working life of a significant researcher."(Corti & Backhouse, 2005)

The Potential for Redistribution and **Full Documentation** criteria assess the formal characteristics of data, format, metadata standards, completeness of documentation, provenance information etc. These criteria may be challenging for NNfD. The variety of formats and standards associated can demand particular technological measures. Data type registries are being considered to support machine-actionable processing (Broeder & Lannom, 2014). A consideration about which actors are best able to preserve certain type of data is also relevant e.g. for textual or speech data a national CLARIN data centres may have specific tools and processes[45]. Legal and ethical characteristics needs to be assessed and clear guidance offered regarding the conditions that the data needs to meet. Consent, anonymity, IPR, etc. all are relevant to the potential for redistribution.

https://www.ukdataservice.ac.uk/media/398725/cd227-collectionsdevelopmentpolicy.pdf

- CURCAT1: Data collections selected for long-term curation.
- CURCAT2: Data collections selected for "short-term" management.
- CURCAT3: Data collections selected for 'delivery' only,
- CURCAT4: Data collections selected for "discovery" only.
- CURCAT5 relates to preservation-only which falls outside the scope of the UK Data Service Appraisal, and which is handled by the UK Data Archive. Data collections may be moved into higher or lower categories over time if the need arises.

[45] FSD advices Finish researchers to store audio material so; Slovenian CLARIN centre stores added value national collection of Twitter data for a given period.

Replication data and resources. One category of data that funders, research journals and the scientific community are considering is replication data and resources. Such data would ideally be accepted as open data, but legal and ethical issues can set limits on sharing, in particular for sensitive data.

Balancing the benefits to society versus a data subjects rights in the GDPR context can be particularly relevant to administrative data. One general feature of administrative social science data is that it 'may also be large, complex and multi-dimensional'. In such cases of limited access, the provision of scripts of research code can be offered for replication purpose alongside or instead of the data. A similar approach may help overcome the barriers that social media companies' terms of service impose on archiving of full content (Kinder-Kurlanda et al., 2017).

Flexibility of processing is key to considering the appraisal and selection criteria for NNfD. As the data can be assessed on the different dimensions, so additional processing steps should be aligned with specific combinations of data characteristics.

Critical outcomes of an appraisal and selection decision include:

- Time scale (for how long will the data be retained/preserved)
- Access conditions (following the legal and ethical characteristics of data, ranging from open to restrict or no access).
- Preservation and curation strategies that affect the end products' long-term usability (levels and aspects of quality checks and control and added value to data; formats, metadata, etc.). If data is enriched and quality controlled, e.g. geolocations added, cleaned data, standard ontologies used, etc. it can be made more accessible to variety of users. Quality controlled data (collections that are important due to their coverage, added value, format, semantic or temporal dimension), that are collected with the idea that it builds a resource for other, with a purpose defined broadly, which a lot of resources have been invested in, are in principle valuable assets that deserve long-term curation. The question is if we can identify such type of resources in the NNfD environment? (Kondyli et al., 2017) [46] Multidisciplinary perspective and generalised criteria should be sought in the EOSC framework, which requires collaboration among preservation and access infrastructures. A common approach to costing and consistent FAIR data specifications can assist this goal.
- Access interfaces (specific tools for searching, linking, documenting, analysing data)

Conclusions & Recommendations

In designing an appraisal and selection process for new and novel forms of data (NNfD) supported by existing and emerging technologies within a mixed ecology of traditional and novel research opportunities we can call on a range of archival approaches. Some traditional approaches to quantitative data can be applied directly to NNfD. Archives can also apply qualitative data appraisal and selection criteria designed to deal with more heterogeneous data types and sources. Disciplinary archives' experience in evaluating the potential value alongside the potential costs of curation of the long-tail of research data can also be applied as less familiar forms of data are offered for deposit. Assessments based on these criteria sits alongside decisions about whether the data reflect the archival mission and align with the established data collection policy. Assessments must be made as to whether the repository can provide the expertise and tools required to support curation, and the overall cost of delivering a service around the data must be estimated.

[46] See SaW results D3.5. about CESSDA current coverage of Big data related to specific domains: "Big data is an all-encompassing field, in the sense that each domain now deals with this new format of data." See http://cessdasaw.eu/content/uploads/2017/11/D3.5_CESSDA_SaW_v1.3.pdf

A number of relevant evaluation criteria are presented below:

Scientific relevance:

- Content
    - reach, new aspects of phenomena, new methods, potential for new analysis
- Historical value
    - Unique perspective, Interdisciplinary value, cultural heritage
- Methodological quality
    - sampling, measurement, combination of different sources, coherence
- Passive vs. active observation
- Observations with time or geographic dimensions
    - longitudinal research, time series, regional comparison
- International comparative character
- Added value product
- Multiple sources combined
- Other services exist for the same type of data/ risk of loss

Replication and validation of research

Educational value

- disciplinary tradition, new methods learning material

Legal and ethical consistency:

- Commercial terms and conditions (ToS)
- Commercial character, value
- Intellectual Property Rights (IPR) clarity
- Anonymity
- Public/private (perception)
- Reporting about ethical and legal concerns while reporting results
- Consent type/ personal data risk assessment
- Purpose limitations
- Covert observation

Formal aspects

- Size
- Format
- Periodicity of updates
- Versioning, citation possible
- Documentation available
- Metadata, FAIR-ness
- Structure
- Complexity

The criteria above may all be evaluated as general data characteristics for appraisal and selection. The presence, absence or degree of each criterion may be integrated into a Collection Development policy as a reason to accept or reject data for deposit.

In the context of survey, administrative and social media data the key, high-level characteristic remains whether the data, or the sources from which it derives, contain personal data i.e. whether the data are personal data, whether they are anonymised data with a risk of re-identification (particularly through data linkage services provided by the repository), and the sensitivity of the personal data.

A repository that does not have the expertise or information security provisions to handle personal data will see this factor as a key decision driver. But for environments which do accept personal data those deposits which include special categories of personal data or which include vulnerable populations will require an additional analysis to compare the risk of data curation and distribution to the overall value of the data to users.

A Collection Development Policy seeks to provide guidance on the acceptance or rejection of data offered for deposit to the archive at the current time. But it is equally important to plan for change.

The increased demand for data to deliver impact beyond pure, traditional academic research, including to policy makers, the media and citizen scientists, requires that archives consider whether their current 'designated communities' must be widened to include these groups. If so there will be a need to monitor the knowledge base and technical needs of this wider group.

For NNfD a key consideration is the sustainability of expected data quality and service offerings in the face of new and complex formats or vastly increased data sizes and throughputs. It may be sufficient during appraisal and selection to reject data which can be stored in alternate environments with basic metadata and bit-level integrity assured until further demand is identified. But given the low rates of NNfD deposit to date it may be wise for repositories to take a precautionary approach to accepting data for which demand has not yet been identified. Only by developing further experience with these data and through ongoing interaction with data providers and user communities can archives monitor demand and consider bringing these data into a full-preservation environment.

In a period of perhaps unprecedented change across data sources, technologies and user communities the appraisal and selection process should be integrated into strategic planning for change. For each case where high quality data of uncertain long-term value is rejected for deposit, or more ideally, accepted for a lower level of curation or redirected to an alternate data steward, the archive must consider what changes would be required to support these data in the future. Options include scaling operations, adopting new infrastructure elements (skills, processes and technologies) or collaboration with other partners.

For many of the criteria identified there will need to be amendments and updates as the research community in general, and archives and other trusted digital repositories in particular, gain more experience with NNfD.

The legislative pressure of GDPR applies to those curating personal data, but we can reasonably expect these drivers for increased transparency of processing to influence future process and technology design which can be applied to a wider range of data sources. Clear processing practices, custody information and other data provenance developed for personal data is of value to all data stewardship workflows.

Administrative, survey data and social media data generators initial focus is on their primary governmental, academic or commercial audience. In data and technology development the priority is to present a product to the consumer; the underlying documentation of structure, function and use may be a secondary consideration. As standards, legislation and best practices are adopted and as demand for this information increases (e.g. when partners need to adopt similar approaches to manage data) there is a drive for initial documentation, formal public documentation and eventually for the development of standardised structured metadata. Progress towards this ideal structured metadata situation can vary but should be supported and the costs of documentation and standards compliance built into research data management plans as early as possible. This not only reduces 'downstream' curation costs but also provides a more robust, FAIR and trustworthy ecosystem throughout the (research) data lifecycle.

The European Open Science Cloud (EOSC) is one example of work to consolidate infrastructure and services which will mean that more data stewards are involved in wider partnerships, with consequent increased numbers of data copies, dataflows and custody transfers, both in cloud and on-premises infrastructures. The emergence of new datastore technologies to support access to linked open data (LOD) and new data analytic paradigms, particularly machine learning and AI, will exist alongside more traditional file and database-driven technical systems for the foreseeable future. Archives must address and manage these increasingly complex partnerships and hybrid technical environments. As consolidated infrastructure offers centralised key functions as services - such as the provision of format registries and researcher identifiers - we must consistently describe and manage a large number of remote dependant objects.

The EC Recommendation on access to and preservation of scientific information reiterates a commitments to these issues and presents a mechanism to support international progress but this must be undertaken alongside the rapid adoption and evolution of FAIR data standards and the European Open Science cloud. Established repository standards and a long history of rapid response to changes in data, technology and user needs put the archives in a good position to support this evolution, but limited NNfD deposits with a focus on long term preservation to date means these approaches have yet to be fully tested in the real world. Measures must be taken to ensure an unprecedented period of data generation and opportunity is not to become an unprecedented period of data loss.

All appraisal and selection decisions must consider the potential costs of maintaining data, for some NNfD there is a balance to be struck in preserving of uncertain future value as a precautionary measure. Along with the full range of academic, scientific and other data actors the key requirements are to maintain centres of excellences capable of managing the transition to long-term preservation of new and novel forms of data offering new opportunities for research using new forms of technology. One critical success factor will be continued professional development and training of data collectors, curators and users, ideally aligned with increased public data literacy and understanding of the societal value research data can bring. The quid pro quo is that scientific infrastructures must develop, document and demonstrate appropriate technical and organisational measures to support Trusted, FAIR and transparent good practices across their data processing activities.

Appendices

Appendices: EOSC, FAIR data, Preservation and the GDPR

Common Appendices for SERISS Project Deliverables 6.8 Versioning requirements for curation and access to new forms of data and 6.9 Appraisal/Selection Requirements for New Forms of Data

Appendix A: Stakeholder Ecosystem- Snapshot

There are several emerging plans, proposals and actions relevant to the European Science Ecosystem that are relevant to planning for new and novel forms of data (NNfD) infrastructure (people, processes and technology) in general and to the work of SERISS T6.3 in particular (workflows, versions, appraisal & selection, metadata)

Among these emerging factors are the revised Recommendation on Access to and Preservation of Scientific Information (EC, 2018), the Implementation Roadmap for the European Open Science Cloud (EC, 2018) and the Turning Fair Data into Reality interim report (Hodson et al 2018), while in the legal sphere the General Data Protection Regulation (GPDR, EU 2016) is now in place. This work takes account of these new development, acknowledging that these are both key concrete factors to be addressed, but also emerging areas subject to further development and change.

The EC Communication "European Cloud Initiative" published on 19 April 2016 has three pillars:

- EOSC: the European Open Science Cloud
- EDI: the European Data Infrastructure: a High Performance Computing (HPC), data and network infrastructure
- Widening access & building trust (Going beyond research to include government, industry and small/medium enterprises (SMEs)

The EOSC Declaration & Implementation Roadmap

The EOSC declaration released in October 2017 (See Appendix C for a summary with a repository focus) provides the initial framing of the ecosystem within which repositories must evolve and function for the medium to long term.

The declaration is inclusive and ambitious from the start, stating that “Only a considerable cultural change will enable long-term reuse for science and for innovation of data created by research activities: no disciplines, institutions or countries must be left behind”. In addition to **data culture** the key topics covered which impact repository actors are:

Open access by default, the evolution and professionalisation of **data stewardship skills**, the development and adoption of **standards** including “technical, semantic, legal and organisational”. **FAIR** is fully integrated through a set of **implementation/transition**, **data governance** and **accreditation/certification** goals. **Data Management Plans (DMP**) are central to the pre-repository phase ad **technical implementation** incorporates the need for **citation systems**, **common catalogues**, a **semantic layer** and other **FAIR tools and services.**

The declaration envisages an implementation roadmap supporting the move towards:

- Clear governance framework
- Definition of the initial services
- Clear business model
- Cost optimisation

The roadmap (published in March 2018) as a commission staff working document integrates a number of these key topics and goals and aligns them with current project and funding activities.

The objective of the European Open Science Cloud (EOSC) is to offer ‘1.7 million European researchers and 70 million professionals in science and technology a virtual environment with free at the point of use, open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines’ (Implementation Roadmap for the EOSC, 2018, p3) which will, through the reduction of fragmentation through a federated research infrastructure offer “every European researcher the possibility to access and reuse all publicly funded research data in Europe, across disciplines and borders” (ibid) acknowledging that “EOSC would need to be both scalable and flexible, adaptable to the emerging needs of the scientific community and able to support the whole research data lifecycle.”

Critical to the repository context is that “funders are gradually tying their funding to open access obligations and the use of FAIR accredited/certified repositories”

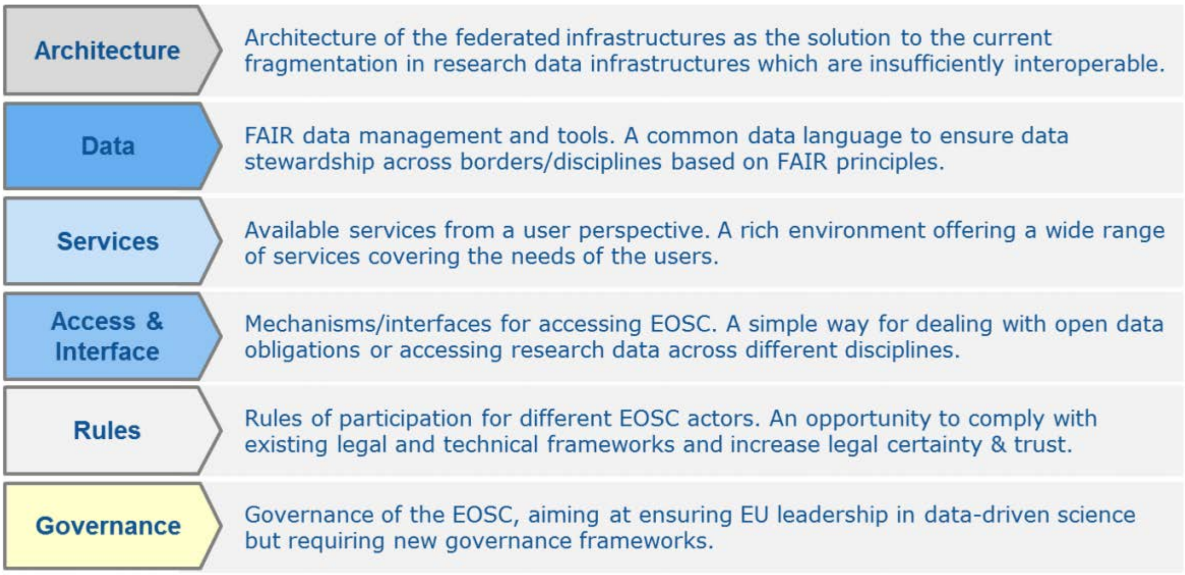
Under the data remit it is made clear that specific action need to be taken:

- to develop a better culture of research data management and practical skills among EU scientists and innovators, including action on incentives, rewards, skills and curricula related to research data and data science;
- to develop FAIR data tools, specifications, catalogues and standards, and supply-side services to support scientists and innovators, and
- to stimulate the demand for FAIR data through consistent FAIR data mandates and incentives to open data by research funders and institutions across Europe.

The envisaged services are:

1. A unique identification and authentication service and an access point and routing system towards the resources of the EOSC.
2. A protected and personalised work environment/space (e.g. logbook, settings, compliance record and pending issues).
3. Access to relevant service information (status of the EOSC, list of federated data infrastructures, policy-related information, description of the compliance framework) and to specific guidelines (how to make data FAIR, to certify a repository or service, to procure joint services).
4. Services to find, access, re-use and analyse research data generated by others, accessible through appropriate catalogues of datasets and data services (e.g. analytics, fusion, mining, processing).
5. Services to make their own data FAIR, to store them and ensure long-term preservation.

In addition to these the EOSC model action lines cover access/interfaces, rules of participation and governance as outlined in the diagram below



Architecture	Architecture of the federated infrastructures as the solution to the current fragmentation in research data infrastructures which are insufficiently interoperable.
Data	FAIR data management and tools. A common data language to ensure data stewardship across borders/disciplines based on FAIR principles.
Services	Available services from a user perspective. A rich environment offering a wide range of services covering the needs of the users.
Access & Interface	Mechanisms/interfaces for accessing EOSC. A simple way for dealing with open data obligations or accessing research data across different disciplines.
Rules	Rules of participation for different EOSC actors. An opportunity to comply with existing legal and technical frameworks and increase legal certainty & trust.
Governance	Governance of the EOSC, aiming at ensuring EU leadership in data-driven science but requiring new governance frameworks.

Figure 1: EOSC Model Action Lines

The roadmap further described planned integration with the European Data Infrastructure (EDI) through provision of high-bandwidth networks and the supercomputing capacity necessary to access and process large datasets stored in the EOSC via the EuroHPC Joint Undertaking (JU).

The main text concludes noting that “in principle, the business models and funding streams of existing data infrastructures should not be affected by the development and operation of the EOSC, as long as they are compatible with the operating principles of the EOSC” (ibid, p21), but also made it clear that:

“little to no published information exists on the current level of spending on research data infrastructures and FAIR data management in Member State and this, along with the variable situation across the EU, is why it is not possible to attach concrete figures to these costs consistently across EU28. The Final report of the High Level Expert Group on the EOSC estimated that on average about 5% of total research expenditure should be spent on properly managing and stewarding data in an integrated fashion.”

These expectations and uncertain costs will need to be incorporated into future repository planning, including selection, appraisal and collections development.

NB: Annex 2 of the roadmap provides succinct descriptions of a number of related entities and actions: EOSCPilot, EOSC-Hub, OpenAIRE-Advance, Freya, eInfraCentral, RDA Europe 4.0, GEANT, HNSciCloud

Recommendation on Access and Preservation

See Appendix B for additional detail on this recommendation.

The Recommendation on access to and preservation of scientific information Brussels, 25.4.2018 C(2018) 2375 final released in April 2018 provides an updated vision from the EC on Access and Preservation (see Appendix B) which “”recognises that big data and high-performance computing are changing the way research is performed and knowledge is shared, as part of a transition towards a more efficient and responsive open science". The recommendation provides both the mechanism and the context within which archives must progress to meet these changes with each member state providing a contact point to support "better definitions of common principles and standards, implementation measures and new ways of disseminating and sharing research results in the European Research Area" for "all research outputs from all phases of the research life cycle (data, publications, software, methods, protocols, etc.)". Recital 7 states:

> *"Preservation of scientific research results is in the public interest. [...] Mechanisms, infrastructures and software solutions should be in place to enable long term preservation of research results in digital form. Sustainable funding for preservation is crucial as curation costs for digitised content are still relatively high. Given the importance of preservation for the future use of research results, the establishment or reinforcement of policies in this area should be recommended to Member States."*

The federation of existing data infrastructures into the EOSC is envisaged through concrete objective, progress indicators, implementation plans and financial planning across the key areas of focus:

- Open access to scientific publications
- Management of research data, including open access
- Preservation and re-use of scientific information
- Infrastructures for open science
- Skills and competences
- Incentives and rewards

For preservation and re-use of scientific information the recommendation focuses on effective deposit systems, ensuring that scientific information selected for long term preservation "receives appropriate curation, along with hardware and software necessary to allow the re-use", and that conditions permit "value added services" based on re-use. Persistent unique identification for findability, reproducibility and preservation is covered within the wider context of linking "research outputs, researchers, their affiliations and funders, and contributors" and there is a clear intention that licensing systems and conditions should become machine-readable.

Prompting an EOSC in Practice

The interim report and recommendations from the High Level Expert Group (HLEG) on the EOSC[47] demonstrates the current challenges for repositories in seeking to standardise their approach to handling NNfD. Released in mid-June 2018 this interim report indicates that the final version will define the features of an EOSC 'minimum viable ecosystem' (p20) and outlines initial thinking on Rules of Participation for federating existing infrastructures (p28) including consideration of private sector users stating "By participating, private sector may want to invest in the long term development and sustainability of the EOSC, along with the public sector and not just serve to exploit public data for free".

Rules of participation cover capacity (computational, storage and network), accessibility ("minimum set of interfaces for data deposit and download, as well as capabilities to launch analytic tools against data deposited at the site"), Identifiers/Metadata ("to understand the data, software or workflow that is being evaluated for reuse") and Information Assurance and Protection by design (including a shared security model, data protection by design, data minimisation and protection of subjects' rights)

Turning Fair Data into Reality

The integration of the FAIR data requirements into organisational and digital object management is presented in the interim report from the European Commission Expert Group on FAIR data: Turning FAIR data into reality (Hodson et al, 2018).[48] The report makes a

[47] https://ec.europa.eu/info/sites/info/files/conferences/eosc_summit_2018/prompting_an_eosc_in_practice_eosc_hleg_interim_report.pdf

[48] https://zenodo.org/record/1285272

clear early statement that the concepts of Findable, Accessible, Interoperable and Re-Usable are an excellent distillation of the mission and goals, but do not provide the best top-tier from which to consider implementation. A "holistic and systemic approach and to describe the broader range of changes required to achieve FAIR data" (ibid, p3) is taken. The report consists of 31 Metrics with Rec 1-14 incorporated into the executive summary. It has not been possible to fully analyse the implications due the recency of release, but this report and its successor will impact repository approaches to workflows and digital object management, including approaches to NNfD.

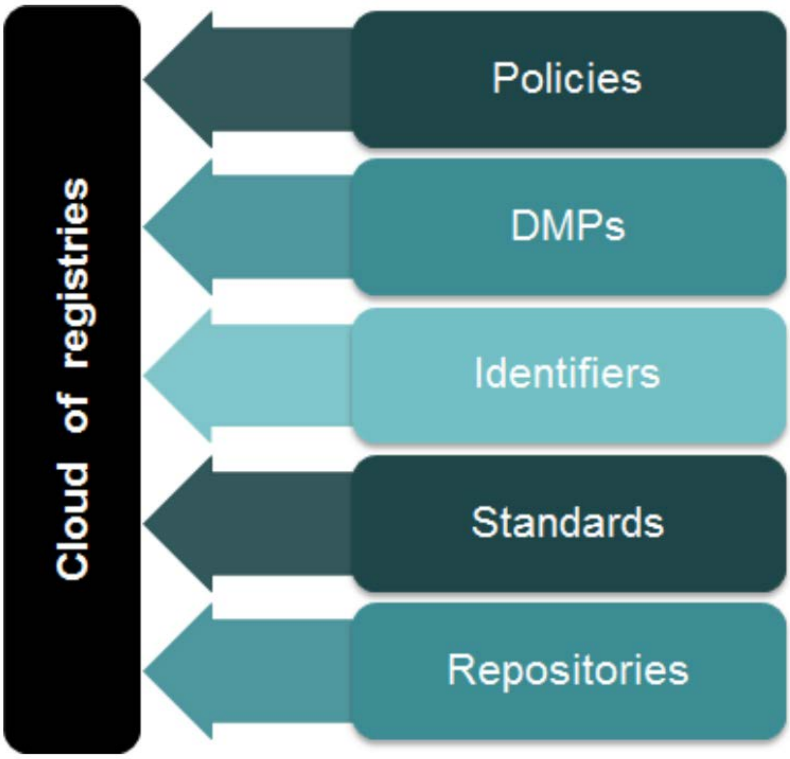


Figure 2: Components of a FAIR data ecosystem

Five components of a FAIR data ecosystem are presented with policies, data management plans, identifiers, standard and repositories all feeding into a 'cloud of registries'.

For the purposes of this paper we consider the initial 14 recommendations, highlighting in bold the terms of most relevance to operationalising these goals in repositories. Recommendation numbers are presented in square brackets:

Step1: A clear definition of FAIR extended to incorporate **openness**, **accessibility** and **long-term stewardship** [1] supported by **mandates** for open data with appropriate **boundaries** (as open as possible, as closed as necessary [2]. FAIR **object models including metadata (PID, provenance and licencing**) supported using **open standards** and **shared code** [3].

Step2: Critical **infrastructure components: policies, DMP, identifiers, standards and repositories**, supported by **automated workflows** and **registries** [4]. With components **professionally** and **sustainably maintained** [5] through **strategic funding** based on evidence (**impact, adoption, certification**).

Step 3**: Disciplinary interoperability** [8] through **common standards, intelligent crosswalks, brokering mechanisms** and **machine learning,** with frameworks incorporating principles for **sharing, agreements, formats, standards tools and infrastructure** [7].

Step 3 (continued): robust, managed **FAIR metrics** for **data objects** [9] stored in **CoreTrustSeal Trusted Digital Repositories** [10], surrounded by **sustainable**, **managed**, **certified data services** [11].

Step 4: regularly updated **Data Management Plans** as **information hubs** for FAIR digital objects [12]. **Professionalised Data Science** and **Data Steward** roles [13] **recognised** and **rewarded** (alongside **infrastructure** and **services**) for **FAIR object management** and **curation** [14]**.**

Appendix B: The EC Vision of our Access and Preservation Mission

This section is a revised version of a post published in cooperation with the Digital Preservation Coalition (DPC)[49].

On the 25th of April this year (2018) the European Commission released its *Recommendation on access to and preservation of scientific information* Brussels, 25.4.2018 C(2018) 2375 final. This work by Mariya Gabriel and Carlos Moedas replaces that by Neelie Kroes (then Commission Vice-President) back in 2012. This revision ”recognises that big data and high-performance computing are changing the way research is performed and knowledge is shared, as part of a transition towards a more efficient and responsive open science".

[Squared brackets] below refer to the 12 recommendations while the document itself is divided into eight unnumbered sections which are italicised below. The opening 15 recitals are identified with (parentheses)

Mechanisms and Reporting

[12.] Envisions member states reporting their actions to the Commission in eighteen months, then every two years thereafter. A report in 2015 provides an overview by member states of progress against the 2012 recommendation.

Structured coordination of Member States at Union level and follow-up to this Recommendation identifies a need for each member state to coordinate the recommended measures and provide a contact point to the Commission towards "better definitions of common principles and standards, implementation measures and new ways of disseminating and sharing research results in the European Research Area" [11.]. These contact points will ensure *multi-stakeholder dialogue on open science at national, European and international level* with an explicit expectation of systemic, gradual changes in research culture across the relevant actors covering "all research outputs from all phases of the research life cycle (data, publications, software, methods, protocols, etc.)"[10.]

Recitals

The opening 15 recitals provide context, particularly around the Digital Single Market Strategy and the European Cloud Initiative but the critical statement for a digital preservation audience may be (7):

> *"Preservation of scientific research results is in the public interest. [...] Mechanisms, infrastructures and software solutions should be in place to enable long term preservation of research results in digital form. Sustainable funding for preservation is crucial as curation costs for digitised content are still relatively high. Given the importance of preservation for the future use of research results, the establishment or reinforcement of policies in this area should be recommended to Member States."*

(7) also notes that this has *traditionally* been the remit of archives and libraries, though we would expect that the vision of the European Open Science Cloud (EOSC) as a "trusted, open environment for the scientific community for storing, sharing and re-using scientific data and results" is seen as a common infrastructure component rather than a replacement

[49] https://www.dpconline.org/blog/ec-vision-of-access-and-preservation-mission

for the 'traditional' centres of preservation practice... The quoted statement from the European Cloud Initiative states that this will "start by federating existing scientific data infrastructures, today scattered across disciplines and Member States" (8).

The remainder of the recitals plays to the requirement that "all accessible data held by a public sector body needs to also be reusable for commercial and non-commercial purposes by all interested parties under non-discriminatory conditions for comparable categories of re-use and at the marginal cost linked to the distribution of the data, at maximum" (Directive 2003/98/EC) (4)) with a scattering of the key words one might expect: open access, use and re-use, licensing, collaborative, volume, professional development.

Though the mechanisms are European, there is an acknowledgement that this is a "worldwide endeavour" with a need for a response on a "global level" (12).

Recommendations

The main recommendations [1.] to [9.] each state a clear requirement for national action plans and policies which provide for:

- concrete objectives and indicators to measure progress;
- implementation plans, including the allocation of responsibilities and appropriate licensing;
- associated financial planning.

Open access to scientific publications

Is explicit about the stakeholders who should be able to access scientific publications: "innovative companies, in particular small and medium-sized enterprises, independent researchers (for instance citizen scientists), the public sector, the press and citizens at large" [1.]. With a clear requirement for transparency about agreements between public institutions and publishers and a target for all publicly funded research publications to be open access by 2020 and that these become open no later than six months after publication (twelve months for social sciences and humanities). This vision of openness includes licence terms which "do not unduly restrict text and data mining of publications".

Delivery is to be supported by institutional policies, guidance on compliance, funding for dissemination and open access as a condition of funding [2.].

Management of research data, including open access

Calls for data management as standard from the point of data collection or generation, with information "as open as possible as closed as necessary", FAIR compliant (findable, accessible, interoperable and re-usable), and held within a "secure and trusted environment" [3.]. Access to and preservation of research data is to be assured through data management planning skills and digital infrastructures (including EOSC) with an explicit requirement that "datasets are easily identifiable through persistent identifiers and can be linked to other datasets and publications through appropriate mechanisms, and that additional information is provided to enable their proper evaluation and use".

Stakeholders are as [1.] above and delivery reflects the same targets of policy, guidance and funding support. A national requirement for DM plans is mentioned, as is their inclusion as a basic principle in grant agreements and other financial support.

Preservation and re-use of scientific information

[5.] is brief and to the point, recommending preservation policies, effective deposit systems, ensuring that scientific information selected for long term preservation “receives appropriate curation, along with hardware and software necessary to allow the re-use”, and that conditions permit “value added services” based on re-use.

Persistent unique identification for findability, reproducibility and preservation is covered within the wider context of linking “research outputs, researchers, their affiliations and funders, and contributors” and there is welcome guidance that licensing systems and conditions should become machine-readable.

Infrastructures for open science

[6.] and [7.] reflect a drive towards researcher access to "resources and services for storing, managing, analysing, sharing, and re-using scientific information" through economically efficient infrastructures. The quality, reliability and interoperability of these infrastructures (including EOSC) are to be assured through data and service standards and metrics that support evaluation of research, careers, impact and openness.

Skills and competences

[8.] Seeks continuous relevant training throughout education and work covering: open access, data research management, data stewardship, data preservation, data curation and open science. With specific reference to data specialists, technicians and data managers in data-intensive computational science.

Incentives and rewards

Evaluation of researcher recruitment, careers, research grant award processes and research institutions are all in scope here. Support and rewards are sought for early sharing and open access to publications and other research outputs with a clear focus on 'new generation metrics' that also provide indicators about the "broader social impact of research" [9.].

From a repositories and archives perspective the recommendations provide concise criteria as we progress towards integrated infrastructures and add linked open data at scale to the file-driven technologies we’re familiar with. It might even provide some guidance as to the ’technical and organisational measures' for which data stewards must provide evidence under the GDPR.

Appendix C: The EOSC Declaration: Summary with a Repository Focus

This outcome of the EOSC Summit of 12 June 2017 was released in late October 2017. Square brackets are key items in the text.

Data culture and FAIR data

[Data culture] “European science must be grounded in a common culture of data stewardship, so that research data is recognised as a significant output of research and is appropriately curated throughout and after the period conducting the research. Only a considerable cultural change will enable long-term reuse for science and for innovation of data created by research activities: no disciplines, institutions or countries must be left behind.”

[Open access by-default] with data 'as open as possible and as closed as necessary' acknowledging that additional protection is required for “personal data protection, confidentiality, IPR concerns, national security or similar” reasons.

Development of “research data management, data stewardship and data science” [Skills] ensuring the availability of sufficient [Data Stewardship] with [Rewards and Incentives] for openness and FAIRness of data and “data-related algorithms, tools, workflows, protocols, services and other kinds of digital research objects”

[Standards] including “technical, semantic, legal and organisational” with an acknowledgement of variation and domain-specific needs.

[FAIR Data governance] covering policy, technical and human resources, social infrastructure for “well-established frameworks and decision-making flows” ensuring “transparency, representativity and accountability” as we “align [...] data-related business processes, responsibilities and expectations to achieve commonly agreed goals”

[Implementation & transition to FAIR] “in all the phases of data life cycle.”

[Research data repositories] “Trusted research data repositories play a fundamental role in modern science. Scientist must be able to find, re-use, deposit and share data via trusted data repositories that implement FAIR data principles and that ensure long-term sustainability of research data across all disciplines. Data repositories must be easy to find and identify, and provide to users full transparency about their services.”

[Accreditation/certification] “clear rules and criteria” FAIR compliant data and for deposit/access infrastructure through " an accreditation or certification body” of certified repositories. “Experience from existing accreditation processes must be taken into account.”

[Data Management Plans] (DMP) “obligatory in all research projects generating or collecting publicly funded research data, [...] minimum conditions for DMPs must be defined, [assured by] host institutions [and provided to] to data repositories.

[Technical implementation] through the provision of: [Citation system], [Common catalogues], [Semantic layer] to support syntactic and semantic interoperability/exchange and access to [FAIR tools and services]

[Data expert organisations] such as RDA, CODATA and the DDI Alliance provide forums to reach FAIR data consensus at European and Global level.

Research data services and architecture

[EOSC architecture] to federate national and disciplinary resources and services, ensuring sustainability through a “continuous dialogue to build trust and agreements among funders, users and service providers is necessary for sustainability”. [Implementation] through objective criteria and stakeholder driven governance towards federation that provides “core common services, certification activities, joint-procurement initiatives, definition of minimum quality standards of service (based on clear Service Level Agreements SLAs), identity provisioning and management, common cataloguing data and computing/analytic services and tools”. Implementation will re-use [Legacy] (current) local, national and European services, solutions and projects to avoid reinventing the wheel.

[User Needs] key requirements to be identified by data scientists, ICT specialists, IT departments, umbrella associations, community networks for [Service provision] which provides “future-proof [...] cutting-edge cloud based environments at high Technology Readiness Levels (TRLs) with a competitive environment to avoid provider lock-in.

[Service deployment] across the lifecycle to provide software, infrastructure, protocols, methods, incentives, training, services through different deployment models “(e.g. Infrastructure as a Service, Platform as a Service, Software as a Service) to communities at differing levels of maturity.

Fair data uptake across [thematic areas] through federation and co-ordination of open data infrastructures.

[Research infrastructures] “The role of ESFRI and EIROFORUM research infrastructures and organisations in the EOSC will be enhanced, Member States and the European Commission made significant investment; research infrastructures should be 'the steward of the community of standards' and provide scientists with a ramp-up for the utilisation of the EOSC.” With [EU-added value and coordination] providing sustainability by ensuring that policy and technology are aligned with national strategies to avoid duplication.

[High Performance Computing and the EOSC] “a pan-European integrated exascale supercomputing infrastructure [to provide] data-intensive advanced applications and services data access and advanced computing and data management services” to support the EOSC.

Governance and funding

A representative, proportional, accountable, inclusive and transparent [Governance model] to support interdisciplinary trust through a [Governance framework] of institutional, operational and advisory functions. With an [Executive board] and [Coordination structure] to ensure [Long-term sustainability] through coordinated funding and income streams and a view towards [Global aspects] as the federated network expands to include global research partners.

Appendix D: Data, Processes and the GDPR

Context

This material has been prepared by SERISS WP6 and T6.3 with the co-authorship of Scott Summers the Senior Research Data Services Officer at the UK Data Service. Scott managed the issue identification and transition process necessary for the UK Data Service and UK Data Archive to ensure compliance with the GDPR and has written and presented extensively on the topic.

NB: this document does not constitute, and should not be construed as proposing a legal or ethical route for any party or data type. It is a working document to support staff and external project partners in defining, at a generic level, the potential impact of the General Data Protection Regulation (GDPR) on workflows for archiving and research.

This material is intended to provide a baseline overview of some of the key concepts around managing data and workflows under the GDPR. It is not intended to provide specifications or recommendations on how the any researcher, ERIC or Archive should or will approach the GDPR, but it could provide some useful background.

The official legislation page is at:

http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=en

Links below are from a third party source which permits links to more granular parts of the legislation.

Introduction

All workflows undertaken related to the data must be permitted either by the original terms of the **consent**, or through another **processing ground**.

The original purpose for processing must be defined. When collecting data for research purposes the most likely grounds for processing the personal data are by (i) consent or (ii) performance of a task carried out in the public interest or (iii) legitimate interest. After initial processing there may be further derogations for scientific, historical research or statistical purposes. Each data controller and processor must understand the criteria which apply to a particular dataset or project and be able to demonstrate compliance and accountability through appropriate technical and organisational measures.

This text does not seek to define appropriate processing grounds for any particular dataset or research environment,

The Digital Objects

> 1) 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, 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 of that natural person;
>
> http://www.privacy-regulation.eu/en/article-4-definitions-GDPR.htm

Non-personal data, even if *linked* to other non-personal, data is not in scope under the GDPR.

> “The principles of data protection should apply to any information concerning an identified or identifiable natural person.”
>
> http://www.privacy-regulation.eu/en/recital-26-GDPR.htm

The digital objects (data, metadata, documentation) in scope are those which

Contain personal data

Contains special categories of data/data from criminal convictions/offences

- Contain **linked data if any of the linked data is of one of the above types** (even if one of the linked data sources is not personal)

In working with these data, we follow the principles of **transparency** (of our actions and the reasons for those actions), **data minimisation** (data should be adequate, relevant and limited to what is necessary) and ensuring that data are not held, or further used beyond their original processing purpose without legal justification. Where data are further processed for research or archiving purposes and the above conditions are met further processing shall not be considered incompatible with the initial purposes.

Digital objects containing personal data must be subject to appropriate **technical and organisational measures** to **mitigate the risk of disclosure** of that personal data.

Versions of the original objects may be created to mitigate risk during processing and storage.

De-identified versions of the data have their direct identifiers (which point explicitly to a person), and indirect identifiers (which could support identification if coupled with other information) removed.

Pseudonymisation is a de-identification method that involves replacing identifying information in the data with artificial identifiers.

Anonymised data has all such identifiers removed so subjects cannot be identified. The data is no longer personal and is not subject to the GDPR.

> "The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.[ ] This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.”
>
> http://www.privacy-regulation.eu/en/recital-26-GDPR.htm

But describing previously personal data as anonymised can be debatable if there are any circumstances under which subjects could be re-identified.

Anonymised versions of data are perceived as having reduced research value.

Pseudonymised data remains personal data under the GDPR.

> “Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person.”
>
> http://www.privacy-regulation.eu/en/recital-26-GDPR.htm

Whether or not data can be considered to be anonymised is not only a characteristic of the digital object, but also the **context of the data situation/environment**.

> “To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.
>
> The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.”
>
> http://www.privacy-regulation.eu/en/recital-26-GDPR.htm

These judgements must be made based on the current data situation, but must also evolve to take account of changes to the data situation over time. The evaluation is of the risk of **re-identification** of data subjects from the data, whether alone or in conjunction with other data sources.

Workflow Artefacts

Several key information artefacts which influence workflows must be managed across the lifecycle.

Consent: The original consent for processing under which the data subject shared their information. Under recital 33 this permits consent (within ethical boundaries) for “certain areas of scientific research”.

Explicit Consent: required for processing special categories of data/data from criminal convictions/offences unless Art 9(2)(j) is used.

Deposit Rights Management: defining the permissions, prohibitions and duties (and any constraints on these) under which the data is held by the archive. E.g. through a deposit licence or agreement.

User Rights Management: defining the permissions, prohibitions and duties (and any constraints on these) under which the data is used by the end-user. E,g through an end user licence or agreement.

Actors and Roles

A wide variety of actors may be involved with workflows across the data lifecycle, each taking on one or more roles. Roles are in italics below, while other relevant concepts are in bold.

The key role is the *data subject* who shares their information on the understanding that it will only be used within the bounds of the original consent or within some other legal justification. The data subject retains a number of **rights** over their data after it is collected.

The data from each subject will be held in a digital object, or collection of digital objects, which are associated with a defined *data controller*. The data controller determines the purpose and the means of data processing and assigns the role of data processor. The *data processor* processes the data on behalf of the data controller.

One or more actors may undertake the following roles while also being the data controller or data processor:

The *data creator* may be a survey organisation, or other research project.

If the data creators are linking the data they collect with other data sources they will work with a *linking data provider*, which may be a government department holding administrative population data or a social media network holding data streams of user interaction.

The *data collection agency* will interact with the data subject and pass data to the data creator.

The *contact management agency* will be responsible for maintaining ongoing contact with the data subject, including the mediation of any requests after data collection that are based on the data subjects’ rights.

The *access provider* mediates access to the data. This may be undertaken by actors prior to deposit in an archive.

The *depositor* acts as the contact point for the deposit of data into an archive.

The *archive* undertakes to curate (process, enrich) store and provide access to the data.

The archive may also ensure the **long-term preservation** of the data by ensuring it remains accessible to a defined community. Preservation is assured by ensuring the data remain understandable (by updating relevant description and contextual information) and usable (by emulating original environments or forward migrating data for use in modern environments).

The archive **mediates access** to the data within the licence terms.

The archive may **mediate use** of the data through the provision of environments for end-users to interact with the data. These could include safe rooms or secure remote access. The usage environment may allow for **data linkage**. If re-identification of data subjects might be a result of data linkage **statistical disclosure evaluation** may be undertaken to **mitigate risk.**

Context for Actions

The actions which may be undertaken upon personal during processing are guided by the principles under Article 5 ‘Principles relating to processing of personal data’

‘**lawfulness, fairness and transparency'**

Original purpose. Processes within the specified explicit and legitimate purposes for how it was collected. ‘**purpose limitation’** is retained (derogations: **Data minimisation** (adequate, relevant and limited to what is necessary)

Storage limitation. No longer than necessary for original purpose, extensions possible for public interest, scientific or historical research purposes or statistical purposes

Integrity and Confidentiality, *data security*

Demonstrable compliance for accountability (maintenance of records)

Derogations from subjects’ rights

The GDPR permits Member States to make derogations to some of the data subjects’ rights under Article 89.

Research and archiving derogations from the subjects’ rights to:

15. **Access** (unless the data remain identifiable)

16. **Rectification**

18. **Restrict Processing**

21. **Object**

17. **Erasure** (provided by GDPR rather than at Member State level)

Additional archiving derogations from the data subjects’ rights to:

20. **Portability** of data

19. be **informed** of rectification, erasure or restriction

Bibliography

Commission Recommendation (EU) 2018/ 790 - of 25
April 2018 - on access to and preservation of scientific information. (2018, April 25). European Commission.

EOSC Declaration. (2017, October 26). European Commission. Retrieved from https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf

European Cloud Initiative - Building a competitive data and knowledge economy in Europe. (2016, April 19). European Commission.

European Commission. (2018). *Prompting an EOSC in practice: Interim report and recommendations of the Commission 2nd High Level Expert Group [2017-2018] on the European Open Science Cloud (EOSC)*. Retrieved from https://ec.europa.eu/info/sites/info/files/conferences/eosc_summit_2018/prompting_an_eosc_in_practice_eosc_hleg_interim_report.pdf

Hodson, S., Jones, S., Collins, S., Genova, F., Harrower, N., Laaksonen, L., … Wittenburg, P. (2018). Turning Fair Data Into Reality: Interim Report From The European Commission Expert Group On Fair Data. https://doi.org/10.5281/zenodo.1285272

Implementation Roadmap for the European Open Science Cloud. (2018, March 14). European Commission. Retrieved from https://ec.europa.eu/research/openscience/pdf/swd_2018_83_f1_staff_working_paper_en.pdf

L'Hours, H. (2018, May 25). The EC Vision of our Access and Preservation Mission. Retrieved August 20, 2018, from https://www.dpconline.org/blog/ec-vision-of-access-and-preservation-mission

REGULATION (EU) 2016/ 679 OF THE EUROPEAN PARLIAMENT AND OF THE C OUNCIL - of 27 April 2016 - on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/ 46/ EC (General Data Protection Regulation).

References

Broeder, D., & Lannom, L. (2014). Data Type Registries: A Research Data Alliance Working Group. *D-Lib Magazine*, *20*(1/2). https://doi.org/10.1045/january2014-broeder

CoreTrustSeal: Core Trustworthy Data Repositories Requirements v01.00. (2016, November). CoreTrustSeal Board. Retrieved from https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf

Corti, L., & Backhouse, G. (2005). Acquiring Qualitative Data for Secondary Analysis. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *6*(2). Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/459

Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., Lemberger, T., … Clark, T. (2017). A Data Citation Roadmap for Scientific Publishers. https://doi.org/10.1101/100784

Davis-Kean, P., Chambers, R., L., Davidson, L. L., Kleinert, C., Ren, Q., & Tang, S. (2017). Longitudinal Studies Strategic Review. 2017 Report to the Economic and Social Research Counci. ESRC. Retrieved from https://esrc.ukri.org/files/news-events-and-publications/publications/longitudinal-studies-strategic-review-2017/

Driscoll, K., & Walker, S. (2014). Big data, big questions| working within a black box: Transparency in the collection and production of big twitter data. *International Journal of Communication*, *8*, 20.

Fiesler and Proferes, 2018 “*Participant Perceptions of Twitter Research Ethics*” Social Media and Society https://doi.org/10.1177/2056305118763366

Genova, F., & Horstmann, W. (2016, September). Long Tail of Data e-IRG Task Force Report. Retrieved from http://e-irg.eu/documents/10920/238968/LongTailOfData2016.pdf

Georgakopoulos, D., Hornick, M., & Sheth, A. (1995). An overview of workflow management: From process modeling to workflow automation infrastructure. *Distributed and Parallel Databases*, *3*(2), 119–153. https://doi.org/10.1007/BF01277643

Kinder-Kurlanda, K., Weller, K., Zenk-Möltgen, W., Pfeffer, J., & Morstatter, F. (2017). Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. *Big Data & Society*, *4*(2), 205395171773633. https://doi.org/10.1177/2053951717736336

Kondyli, D., Fragoulis, G., Linardis, A., Paton, N., Zemborain, F., Ferreira, P. M., … Štebe, J. (2017, November 16). Deliverable D3.5 Report on the state-of-the-art, obstacles, models and roadmaps for widening the data perimeter of the data services. Retrieved from http://cessdasaw.eu/content/uploads/2017/11/D3.5_CESSDA_SaW_v1.3.pdf

Künn, S. (2015). The challenges of linking survey and administrative data. *IZA World of Labor*. https://doi.org/10.15185/izawol.214

L’Hours, H., Butt, S., Henriksen, G., Krejčí, J., Štebe, J., Myhren, M., … Bell, D. (2018a, January). Generic high-level workflows for the curation of ‘Big Data’ Deliverable 6.7 of the SERISS project funded under the European Union’s Horizon 2020 research and innovation programme GA No: 654221.

L’Hours, H., Butt, S., Henriksen, G., Krejčí, J., Štebe, J., Myhren, M., … Bell, D. (2018b, August). Versioning Requirements for Curation and Access to New Forms of Data. Deliverable 6.7 of the SERISS project funded under the European Union’s Horizon 2020 research and innovation programme GA No: 654221.

Lyon, D. L. (2007). Dealing with Data: Roles, Rights, Responsibilities and Relationships. *DEALING WITH DATA*, 65.

Michener, W. K. (2015). Ten Simple Rules for Creating a Good Data Management Plan. *PLOS Computational Biology*, *11*(10), e1004525. https://doi.org/10.1371/journal.pcbi.1004525

Reference Model for an Open Archival Information System (OAIS). (2012). CCSDS Secretariat. Retrieved from https://public.ccsds.org/pubs/650x0m2.pdf

Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (2010, February). Blue Ribbon Task Force. Retrieved from http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf