Deliverable Number: D6.7

Deliverable Title: Generic high-level workflows for the curation of different forms of 'Big Data'

Work Package: WP6 New forms of data: legal, ethical and quality issues

Deliverable type: Report

Dissemination status: Public

Submitted by: CESSDA (UKDA)

Authors:

Hervé L'Hours CESSDA (UKDA)

Sarah Butt ESS ERIC (City/HQ)

Gry Henriksen CESSDA (NSD)

Jindřich Krejčí CESSDA (CSDA)

Janez Štebe CESSDA (ADP)

Marianne Myhren CESSDA (NSD)

Tom Emery NIDI

Darren Bell CESSDA (UKDA)

Date Submitted: January, 2018

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) 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. Available at: www.seriss.eu/resources/deliverables

Contents

Contents

Deliverable Structure, Content and Development. 5
Introduction 6
Approach 7
Terminology 8
Archival, Trustworthy Digital Repository Standards 9
OAIS Reference Model 9
Trustworthy Repository Standards 10
Identifying the Big and the New and Novel 11
Data Characteristics and Workflow Triggers 14
Stakeholder Ecosystem 15
CESSDA Data Archives 16
Survey Research Infrastructures 18
Outsourcing and Complex Partnership Systems 19
Lifecycle & Workflow Management Factors 19
Consent 20
Digital Objects as Data Assets 21
Emerging evaluations of NNfD for archiving 22
Data Linkage 23
Versions for Curation and Access 24
Quality 24
Technical Infrastructure 25
Study-driven workflows 25
Addressing New and Novel Forms of Data 26
Preservation/Long Term Access 27
People, Roles and Skills 27
Legal and Ethical Context and the GDPR 27
Data Lifecycle Perspective 28
Continuous 29
Recurrent 30
Generic High Level Workflows (for Archives) 32
Pre-Ingest 33
Depositor Contact Management Workflow 35
Depositor Correspondence management workflow 35
Deposit Negotiation Management (Appraisal & Selection) Workflow 35
Ingest 36
Data Curation 38
Data/Metadata Publishing 38
Text Box: Negotiating Access and Mediating Use 39
Access 39
Registration Workflow 40
Authentication Authorisation Workflow 40
Access Request Workflow 40
Access Mediation Workflow 40
Data Delivery/Mediating Use Workflows 41
Mediating Use 41
Output Request Workflow 41
Output Mediation Workflow 41
Use-Reuse Support 41
References 42

List of Figures

Figure 1: Mechanisms of Consent Behaviour 20
Figure 2: Schematic of the Collection Pipeline (Kinder-Kurlanda et al. 2017) 23
Figure 3: Functional Lifecycle Model. 29

List of Tables

Table 1: Forms of Data with New Research Potential: 13
Table 2: Data Types by Level of Structure 22

Deliverable Structure, Content and Development.

In this section we present the Work Package 6 and Task 6.3 approach to this portion of the SERISS work. In remaining sections, we will avoid Description of Work (DoW), work package, task and deliverable descriptions and identifiers to ensure the final text is as accessible as possible to as wide an audience as possible.

This first deliverable (D6.7) of Task 6.3 (Connected curation and quality) of the SERISS WP 6 (New forms of data: legal, ethical and quality issues) focusses on the description of high level workflows for the curation of new and novel forms of data (NNfD). As these workflows provide the framework around which other elements of WP6 and Task 6.3 work will be benchmarked and evaluated, we will refine, revise, and expand this text for our subsequent deliverables:

- D6.8 Versioning requirements for curation and access to new forms of data (M36)
- D6.9 Appraisal/selection requirements for new forms of data (M36)

All these deliverables include elements of the T6.3 subtasks covering versioning, appraisal and selection, metadata and data linkage.

This integrated approach will permit a focussed process of outreach, feedback and iteration. The timing of subsequent deliverables permits the inclusion of some real world elements of of the General Data Protection Regulation (GDPR), and to make any necessary amendments to the structure and content of the high-level workflows material.

Task 6.3 of SERISS seeks to describe the current status of archival systems in such a way that we can use these descriptions of the ‘as is’ situation as a benchmark for the findings of T6.1 (Legal and Ethical Challenges related to the use of Social Media Data and related data) and T6.2 (Legal, ethical and quality challenges related to the use of Administrative Data) on the likely impact of new and novel forms of data. In this way, we can describe the ‘to be’ situation implied by these data and consider how our systems must evolve to cope with them. The approach will:

- Model and describe the existing system in a clear, coherent way
- Identify and model the desired features for the future of archives

In this way the final outputs will be as up to date as possible at the time of delivery and will provide a clear citable reference point upon which further work in these areas can be built. The focus of subsequent deliverables on Appraisal/Selection and versioning will also be used to analyse specific data types with a view to identifying the characteristics (size, format, identifiable, structural complexity, lack of provenance, broken custody chain) which will trigger archiving workflows.

Introduction

The mission of the Consortium for Social Science Data Archives (CESSDA) ERIC (European Research Infrastructure Consortium) is to "provide a full scale sustainable research infrastructure enabling the research community to conduct high-quality research in the social sciences contributing to the production of effective solutions to the major challenges facing society today and to facilitate teaching and learning in the social sciences."[1]

In common with all other actors in the (research) data lifecycle the CESSDA member archives are facing a step-change in the data landscape as new and novel forms of data (NNfD) and the technologies to support their curation and use, become more prevalent.

The situation for most data archives is: providing preservation and access services to a wide selection of data and metadata curated to continuously evolving standards over a long period of time; a range of legacy data curation and access tools and products; evolving approach to handling the challenges of data variously described as 'big' and 'new and novel[2]

The nature of new and novel forms of data present a number of challenges and opportunities which indicate a move from mediating the distribution of clearly bounded data files to mediating the flexible creation, access and use of complex data products on platforms which natively support linkage between disparate data formats.

The focus of this work is to define high-level workflows for these **archives** in terms of the challenges presented by these forms of data, but we also describe and contextualise the roles of other data lifecycle actors: the **survey research infrastructures** (European Social Survey (ESS ERIC) and the Survey of Health, Ageing and Retirement in Europe (SHARE ERIC)), and other **survey** partners represented on SERISS[3], (Generations and Gender Programme (GGP), European Values Study (EVS) and the WageIndicator Survey).

By comparing these survey-focussed and archival actors we provide a reference point for how these ERICs face similar, and different, challenges depending on the scope of their missions and the phases of the research data lifecycle they are active within.

We present a view on the challenges of NNfD that takes a distinctly archival perspective, rather than the active data collection and linkage perspective of the survey infrastructures. But we also seek to demonstrate the common challenges of data management throughout the lifecycle and to indicate that the experience of the archives is directly applicable to each actor with data stewardship responsibility.

The audience for our outputs across the European research infrastructures is not limited to a narrow range of stakeholders as cooperation between researchers, policy makers, curators and technologists (including vendors) is required to realise the vision of a future research data infrastructure (see Technical Infrastructure).

The key distinction of an archive as a data lifecycle actor is that it must seek to develop scalable workflow solutions to a range of data, requiring a range of technologies to support a range of end user needs. This contrasts with the surveys and survey infrastructures where

[1] https://www.cessda.eu/Consortium/Mission-Vision

[2] For a summary please refer to page 13, box 2 in Entwisle, B. and P. Elias (2013). New data for understanding the human condition: International perspectives, OECD, Paris, France.

[3] https://seriss.eu/who-is-involved/seriss-consortium/european-social-survey-ess/

the range of data addressed may be narrower. An Archive or Survey may develop 'bespoke' curation processes (tool-chains) to experiment with NNfD, but the Archive must scale from the bespoke to a curation systems which performs at scale. We identify that while archives have experimented with the challenges of handling a number of forms of NNfD, with a particular focus on social media and administrative data sources, there are as yet very limited examples where workflows exist to support archiving these data at scale.

We analyse the NNfD landscape to identify that the key challenges to the data landscape are or a type familiar to traditional data archives. Alongside these changes in the data landscape, we are experiencing a generational change in underlying technologies which mean that the issue of workflows must be addressed alongside the issue of evolving technical infrastructure.

This initial version of the high-level workflows provides a framework which will be expanded upon in future revisions to more fully cover the areas of *data versioning* and *appraisal and selection* for specific data types, sources and formats, as well as to integrate legal and ethical challenges as the General Data Protection Regulation (GDPR) is enacted.

Beyond the necessary context provided by a full lifecycle view and a compare/contrast of survey infrastructures to archives, this work focuses on NNfD solely as an input into archives holding data of interest to those working in social science and humanities (SSH) research. In this way the archives provide assurance of well-curated data of clear integrity and provenance suitable for sharing, re-use, and long-term preservation. Archives, alongside other data lifecycle actors, seek to ensure that curation and support is undertaken from a robust methodological and/or standards-based approach.

Approach

We seek to identify an agreed approach to describing the *Data Lifecycle Perspective* and to define the key *Generic High Level Workflows (for Archives)* which apply to both traditional (research study based) and NNfD (not originally conceived with research, or research infrastructure in mind). This generic approach leaves room for archives with a range of sizes, maturities and data collection foci to 'plug in' locally relevant details.

Addressing the archival curation workflow is not sufficient. We need a clear vision of how these workflows fit into the wider (research) data lifecycle, and how NNfD, and the tools to manage and manipulate them, change the traditional relationship between Archival curators and other stages of the lifecycle. In this way the current and future archiving scenarios may both be described.

With this wider vision in place these high level workflows can contribute to future workflow reengineering (Georgakopoulos, Hornick, and Sheth 1995, 120), a necessary element of effective change management and especially important when addressing fundamental changes in technological infrastructure or a move from manual to machine-mediated or automated workflows.

New and novel forms of data will provide for new curation challenges, not least around legal and ethical issues and the V's of Big Data (Volume, Velocity, Variety, Veracity and Value), but existing methodologies and best practices will be used to address these issues wherever possible. Change management from traditional to NNfD curation is not only a question of evolving data types. We identify that there are certain levels 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.

Archival data services are evolving past offering 'downloads' to involvement in the 'use' phase, particularly with regard to personal data or data with a risk of disclosure. Activities include provision of data linkage, analytics, visualisation functionality, and statistical disclosure control before data products can leave secure access environments.

Note also that 'data' and the objects we use to describe and manage them influence both metadata approaches and the other information 'artefacts' (procedural and evidential) necessary to support new workflows.

Some common understanding of these concepts across governance, management, operational and technical teams is critical to legal, ethical, standards-compliant service delivery and managed change.

Terminology

Terminology is selected to be accessible to as wide an audience as possible. Some degree of specialist methodological and standards-based terminology is unavoidable, and these are defined in situ wherever practical.

Any individual or organisation with responsibility for data storage, curation or associated services during the data lifecycle is a *data steward*.

The *architectural* model designed for use here is based on the continuous, recurrent and sequential activities necessary to support archival data services across the data lifecycle.

These generic *architectural* models which underlay the data lifecycle and curation workflows can be mapped to a wide range of real world archives. An *infrastructure* is conceived as the range of specific skills, workflows and technologies by which we implement the generic architecture activities to support a specific mission around data with clearly defined characteristics. An *ERIC* is one such infrastructure. *Technical Infrastructure* is used to refer to the system, tools and code elements which support the workflows and human actors.

Data may change custody through a complex network of storage or workflow nodes throughout its lifecycle. Any organisation or entity which stores and provides access to data may be considered as a *repository*.

The CESSDA-ERIC member partners are all described as '*archives'* in the OAIS sense (see Standards), these partners are also committed to becoming Trustworthy Digital Repositories (TDR) but the notion of a TDR is not limited to data stewards which self-describe as Archives.

The sequential parts of the archival stage of the lifecycle are referred to as *functions* (OAIS) and each *workflow* is a set of data flows and associated data management activities. Generic high-level workflows can be used to derive more detailed or locally applicable business processes. Business Processes tend to be more detailed descriptions of actions taken including inputs, outputs and roles undertaken by defined actors.

Archival, Trustworthy Digital Repository Standards

To maximise trust, security and impact we seek to maximise the number of trustworthy nodes in the lifecycle, whose consistency and fitness for purpose is the responsibility of trustworthy 'data stewards'

It is insufficient to ensure that data are stored in (and made available from) an environment which supports bit-level integrity checks, multi-copy/multi-site redundancy and low-risk disaster recovery methods. Though these measures are critical, they do not by themselves ensure the full, long-term value of the data assets. Storage providers that ensure bit-level integrity and provide access to data are vital nodes in the data and research data networks, but Trustworthy Digital Repositories are defined by the results they produce i.e. ensuring the availability of data which is understandable and usable by their designated community for the long term.

The high-level workflow elements applied here are derived from a mature set of archival/repository standards. These standards provide important context in the differentiation of Survey/Survey ERIC actors and Archive actors below.

Traditional archival workflows are aligned with these standards and any change to workflows to handle new and novel forms of data, or changes to underlying technical infrastructure, must be managed in a way which ensures continuity of standards-compliance.

OAIS Reference Model

> *Designated Community (OAIS): "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."*

The mandatory responsibilities for an organisation seeking to provide long-term digital preservation are defined by the OAIS reference model ("Reference Mode for an Open Archival Information System (OAIS)" 2012) which makes it clear that an archive shall:

> *"– Obtain sufficient control of the information provided to the level needed to ensure Long Term Preservation.*
>
> *– Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided, thereby defining its Knowledge Base.*
>
> *– Ensure that the information to be preserved is Independently Understandable to the Designated Community. In particular, the Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information."*

(OAIS, 2012)

Clarity on which data are curated for long term preservation, and which are not, is critical to ensuring the persistence of integrity and quality through the data lifecycle, and over time.

Even if data is explicitly *not* intended to be maintained for the long term, the OAIS criteria remain critical to ensuring access beyond the next round of change in the short to medium term, for the longer term a focus beyond digital object management is necessary.

An OAIS accepts one or more deposits from a data *producer* which are defined as a *submission* (Submission Information Package-SIP), these are curated through *data*

management systems. Copies suitable for long term preservation (Archival Information Packages-AIP) are held in *archival storage* and consumers are provided *access* to copies (dissemination information package-DIP) whose format, content and context is appropriate to the designated community. *Preservation planning* ensures that the data remain suitable for preservation and use in the face of changes to technology or changes in the needs of the designated community.

The OAIS provides a reference model, and indicates required elements for baseline archival architecture, but it is not an implementation standard and does not specify infrastructure requirements. The standard identifies the need for criteria against which an OAIS can be evaluated for compliance and a number of these Trustworthy Digital Repository standards now exist.

Trustworthy Repository Standards

> The concept of "circle of trust" is based on the agreement that each member of the circle is accepted according to the same rules and conditions that are approved by all members. In the context of statistical activities, trust involves confidentiality rules and security requirements but also competence and legal aspects.
>
> ("OECD Expert Group for International Collaboration on Microdata Access- Final Report" 2014)

One of the agreed bases for assuring trust in and between CESSDA-ERIC members is certification as a Trustworthy Digital Repository (TDR).

The notion of a TDR stems from the need to move beyond de facto trust in partner organisations to act as responsible stewards of data, towards de jure assertions of their trustworthiness. A progression through standardisation through to audit and certification is common as service models mature. But certification also supports the clear identification of trustworthy 'nodes' in the (research) data lifecycle where, third party relationships (see *Outsourcing and Complex Partnership Systems)* make the infrastructure of data services more opaque.

The notion of trust is critical across the data lifecycle. The 'repository' or 'archive' model has historically defined itself as a distinct part of the lifecycle, but increasingly their disciplinary standards and best practice have been adopted into more general research data management guidance. Organisations which consider themselves as repositories are increasingly 'full lifecycle' actors as they are engaged with data producers pre-deposit and with researchers during the data use phase, particularly for sensitive personal data.

TDR criteria broadly cover the following areas at different depths

- Governance and Organisational Infrastructure
- Digital Object Management
- Infrastructure and Security Risk Management

The least developed of the areas around trusted digital repositories is information security management. ISO16363 repeatedly references ISO27001 for Information Security Management[4] in this regard. For NNfD containing personal information security is

[44] https://www.iso.org/obp/ui/#iso:std:iso-iec:27001:ed-2:v1:en

particularly relevant for workflow and data pipeline management, including in the development of technical and organisational methods to support the GDPR.

To design and change workflows which ensure continued access to, and use of data archives must be able to plan beyond the next round of change to their:

- Designated Community of Users: requiring new data, 'old' data in new formats, new context, new quality information, new access methods etc.
- Technical Infrastructure: moving from extant technology to new technologies that offer the same or new and improved services.

From a lifecycle perspective it is ideal if data quality, integrity and provenance is ensured through stewardship by an unbroken chain of trustworthy actors and actions. From a workflow perspective it is critical to identify which instances or versions of digital objects will persist over time and for how long. This identifies which data will be available for re-use, including to support repeatability and reproducibility.

Identifying the Big and the New and Novel

There are multiple assumptions about, and interpretations of, the concept of 'Big Data'. Big isn't always huge (to the point it presents a challenge to the workflow infrastructure). The key challenges are likely to be associated with the new and novel *sources* of data (with associated new stakeholder relationships, and data provenance and data integrity issues) rather than new and novel *forms* of data.

Adjusting to bigger, faster, more complex data curation and delivery is a standard part of the evolution of data archives' collections. One of the most significant aspects of this 'big data' revolution may be that the technical infrastructure paradigm is changing to go beyond traditional data stored in databases supported by resource discovery and access via web forms. Archives seeking to operationalise 'big data' curation workflows may need to develop or procure access to entirely new technical environments.

The UK ESRC-funded Big Data Network[5] saw the UK Data Service[6] provide support for a range of Big Data Centres including urban big data, business and local government data and consumer data. A key goal was to go beyond the catch-all term of 'Big Data' and the more generally applicable 'New and Novel Forms of Data'. This work has identified that actors in the lifecycle have different perspectives on these data and that, in general, the data themselves (and the challenges they present) are not entirely new to archives. In most cases the challenges are presented by new and novel sources of data and the new and novel 'technologies' required to fully leverage their potential. Workflows implied by this new data landscape are driven by two key features.

- Increased use of personally identifiable data and the need to anonymise or ensure appropriate consent
- The linkage of multiple data sources through new analytical technologies which offer rich research possibilities, but also carry a risk of re-identification of data subjects.

[5] http://www.esrc.ac.uk/research/our-research/big-data-network/big-data-network-phase-2/
[6] https://www.ukdataservice.ac.uk/

This is not to assert that all NNfD are related to personal data, but it is data with this characteristic which present the most profound legal and ethical challenges.

In the report “New Data for Understanding the Human Condition” (OECD, 2013) , there are examples of the wide variety of data that are now becoming more available or are potentially available for research purposes’. Six categories of data are defined in this report:

- Category A: Data stemming from the transactions of government, for example, tax and social security systems.
- Category B: Data describing official registration or licensing requirements.
- Category C: Commercial transactions made by individuals and organisations.
- Category D: Internet data, deriving from search and social networking activities.
- Category E: Tracking data, monitoring the movement of individuals or physical objects subject to movement by humans.
- Category F: Image data, particularly aerial and satellite images but including land-based video images.

In their use of these categories to identify the potential scope of NNfD the UK Data Service concluded there would be a need to address a wide variety of data (not necessarily in high volume) whose key distinction from traditional study collection items was the limited context, provenance and quality in comparison to data generated with research in mind. For all of these categories, but particularly for category D, when these data reach a certain complexity and scale the technical infrastructure must enable intelligent and even autonomous machine-driven processes, discovery, decisions, and applications alongside the human actors. The specific features and challenges of NNfD and the levels of technical infrastructure necessary to support them are covered in subsequent sections.

Broad category of data	Detailed categories	Examples
Category A: Government transactions	Individual tax records	Income tax; tax credits
	Corporate tax records	Corporation tax; sales; tax, value added tax
	Property tax records	Tax on sales of property; tax on value of property
	Social security payments	State pensions; hardship payments; unemployment benefits; child benefits
	Import/export records	Border control records; import/export licensing records
Category B: Government and other registration records	Housing and land use registers	Registers of ownership
	Educational registers	School inspections; pupil results
	Criminal justice registers	Police records; court records
	Social security registers	Registers of eligible persons
	Electoral registers	Voter registration records
	Employment registers	Employer census records: registers of persons joining/leaving employment
	Population registers	Births; marriages; civil unions; deaths; immigration/emigration records; census records
	Health system registers	Personal medical records; hospital records
	Vehicle/driver registers	Driver licence registers; vehicle licence registers
	Membership registers	Political parties; charities; clubs
Category C: Commercial transactions	Store cards	Supermarket loyalty cards
	Customer accounts	Utilities; financial institutions; mobile phone usage
	Other customer records	Product purchases; service agreements
Category D: Internet usage	Search terms	Google©; Bing©; Yahoo© search activity
	Website interactions	Visit statistics; user generated content
	Downloads	Music; films; TV
	Social networks	Facebook©; Twitter©; LinkedIn©
	Blogs; news sites	Reddit
Category E: Tracking data	CCTV images	Security/safety camera recordings
	Traffic sensors	Vehicle tracking records; vehicle movement records
	Mobile phone locations: GPS data	
Category F: Satellite and aerial imagery	Visible light spectrum	Google Earth©
	Night-time visible radiation	Landsat
	Infrared; radar mapping	

Table 1: Forms of Data with New Research Potential:

Like the term 'Big Data' the '4 V's of Big Data' [7] (velocity, volume, variety, and veracity) provide a simplistic but accessible approach to new and novel forms of data. The relevance and importance of each V will depend on the target data, disciplinary context and technical infrastructure, but the fifth V, 'value' remains a key variable when designing high-level workflows for NNfD.

Apart from 'streaming' data most of these new and novel forms of data remain 'snapshots' (timestamped packages) of data taken from a defined universe. Even streaming data is likely to be submitted in snapshots for archival processing.

[7] http://www.ibmbigdatahub.com/infographic/four-vs-big-data

The challenge for an archive developing or amending workflows is to identify the key characteristics of the data under curation. What is truly unique about these data and what is a change of degree rather than of fundamental type? What do archives need to understand about administrative or social media data types to ensure effective curation, access and preservation, or to support linkage?

Archives receive periodic deposits of related data which need to be identified, associated and versioned. However, the velocity might increase to data being generated every 50 milliseconds e.g. network traffic capture. At a threshold of periodicity, the point is reached where current human or machine resources cannot handle data "deposits"; this necessitates a change to technical infrastructure as well as workflows.

Archives have always dealt with increasing data **volumes**. At which point these become 'unmanageable' on current infrastructures depend on the current state of those infrastructures (local organisational context). Volume clearly affects storage capacity but may also impact any data linkage or analysis services on offer – clearly, it takes longer to query a terabyte of statistical data than it does a megabyte.

Archives have always dealt with a **variety** of data. That variety will increase with access to data not created with research in mind, or data that we want to use for different forms of research than originally intended.

Veracity of data sources is an existing issue that becomes more significant with a variety of data sources not originally designed for research. In comparison to study-driven data, these may have drastically reduced documentation of provenance, integrity, chain of custody, past curation actions etc.

From a technical infrastructure perspective the critical differentiators between archives will be whether they maintain a primarily file-based system, the level to which they offer granular access to micro-data through their services (NESSTAR variable browsing[8], Question Bank[9]), and whether they have NNfD-native systems (see Technical Infrastructure) where generically structured data (e.g. RDF[10]) can be addressed and linked across traditional 'object' boundaries.

Data Characteristics and Workflow Triggers

For a Survey ERIC any data with potential for linkage to survey data must be evaluated for their potential to add value or reduce costs. In contrast an Archive must evaluate each potential deposit for alignment with their collections development policy. Both actors must ensure they have the appropriate infrastructure (including technology and skills) to curate the data and linked data appropriately. Once data is within the infrastructure the actor must identify which characteristics of the data impact the choice of workflows.

From a workflow perspective one critical characteristics in this context are whether the data include personally identifiable information or are in an environment where re-identification is a risk. These imply specific legal and ethical measures including risk-mitigation and disclosure control.

[8] http://nesstar.com/help/4.0/webview/getting-started/exploring-data.html
[9] https://discover.ukdataservice.ac.uk/variables
[10] https://www.w3.org/RDF/

Whether actors handling sensitive data for the purposes of research have the option to seek changes to original consent from data subjects is a key workflow trigger. Actors handling only de-identified/anonymised data in environments where re-identification is not a risk avoid many potentially complex and costly workflow requirements.

Stakeholder Ecosystem

> *The broadest understanding of* stakeholders *encompasses any individual or entity which influences, or is influenced by the system in question. The "stakeholder ecosystem" defines the set of entities that we interact with, including users, depositors, funders, legislators, standards bodies, Ethics Review bodies, and employees. Stakeholder identification and management informs the services we deliver."*
>
> *Basic rule 6. Unambiguous distribution of responsibilities should be agreed in advance of any research-related data handling. An important consideration is the potentially complex balance of power between, on the one hand, major public and private stakeholders and, on the other hand, the individuals whose data is being handled."(OECD 2016)*

Repositories/ERICs need to define relevant stakeholders and manage relationships with them. These range from European and National legislators guiding the structure and interpretation of legal requirements to the specific roles in our workflows, including those specified under GDPR (Controllers, Processors, and Data Protection Officers).

Some stakeholder interactions are specific to local organisations; others such as national interpretation of legislation, or case law decisions can best be managed through multi-organisation cooperation. Stakeholder interaction is a 'full lifecycle' planning activity necessary to guide our processes and prepare for change.

The critical subset of stakeholders is those who are direct actors in our workflows, especially when formally defined into business process models (e.g. using UML[11] or BPMN[12]).

Survey and Archival Actors

Below we compare and contrast CESSDA- ERIC Archives, Survey ERICS and other Surveys actors. More detailed actors/roles are addressed alongside the high-level workflows. As we adopt NNfD-native technologies many more activities currently undertaken by human actors will be handled at scale by machine agents.

We consider the full range of responsibilities that any 'data steward' may undertake even if they are not the idealised 'Information Secure, OAIS modelled, Trustworthy Digital Repository'. Each custody transfer between data stewards' presents a risk if these are not 'trusted nodes' in the research data lifecycle and each point of risk must be addressed by the workflows.

> *Research infrastructures (RIs) are facilities, resources and services used by the science community to conduct research and foster innovation.*
>
> *By pooling effort and developing RIs, European countries can achieve excellence in highly-demanding scientific fields and simultaneously build the European*

[11] http://www.uml.org/
[12] http://www.bpmn.org/

Research Area (ERA) and Innovation Union.[13]

They include: major scientific equipment, resources such as collections, archives or scientific data, e-infrastructures such as data and computing systems, and communication networks. RIs can be single-sited (a single resource at a single location), distributed (a network of distributed resources), or virtual (the service is provided electronically).

European Research Infrastructure Consortia (ERIC) are designated by the European Strategy Forum on Research Infrastructures (ESFRI)[14] . The ERICs are instantiated as a wide range of subject, disciplinary and scope initiatives which are not directly comparable from technical or workflow perspectives.

CESSDA Data Archives

Surveys may be broadly categorised as 'full lifecycle actors' (see below) with bespoke technical tooling and workflows for developing and delivering clearly scoped survey data linked to an increasing, but still narrow, range of new and novel (including big) data sources. These survey actors undertake data analysis, but also make data (and linked data products) available to a wider variety of researchers (see *Survey Research Infrastructures* below).

Members of the Consortium of European Social Science Data Archives (CESSDA ERIC) require workflows that support a wider range of data types and formats from a wider range or sources for delivery to a wider range of researchers. Over time the underlying architectural functions of archives (deposit, curation, data management, storage and access) have remained stable while the technical, workflow and human infrastructure have needed to change. Data archives have always needed to ensure continuity of service while managing change in the face of a wide variety of social, scientific and technical issues including new data sources, file formats, quality criteria, resource discovery, and demands for increased data integrity, provenance and granularity. For the archival actors there is a need to maintain information about their workflows, the user communities and their technical environment. This knowledge enables forward planning and change management e.g. identifying the point at which an increased range of data sources, the prevalence of more complex linked data, the scale of data sets and the pace of deposit will start to challenge the technical and human resources available. The history of archives is a history of identifying and adopting new technologies, automating processes for efficiency, and transitioning to new skillsets.

CESSDA Archives deliver a range of data-related services built around the notion of being an OAIS (*Open Archival Information System* ISO14721) which ensures long-term access to (preservation of) digital resources. The OAIS reference model is further supported by audit and certification approaches to designating an organisation as a 'trustworthy digital repository' (TDR) including *Audit and Certification of Trustworthy Digital Repositories Standard* (ISO16363), *Criteria for Trustworthy Digital Archives* (DIN31644), and the CoreTrustSeal (formerly Data Seal of Approval). See Archival, Trustworthy Digital Repository Standards above)

TDR standards provide metrics against which organisational infrastructure (governance, sustainability), digital object management and technical and security infrastructure can be assessed. TDR certification does not preclude the adoption of additional more rigorous or

[13] https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=about

[14] https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=eric-landscape

complementary standards such as ISO27001[15] for Information Security. Deposit of (research) data in an appropriate repository for curation, preservation and access is considered "best practice," and forms a critical part of Data Management Plans (DMP) and research funding body recommendations. Both the OAIS and TDR perspectives include the notion of a 'designated community of users' for whom data must remain usable. Usability takes account of both the knowledge base and the technical facilities available to the community and therefore implies a level of specialist disciplinary and technical knowledge by the repository.

Though CESSDA ERIC Archives are increasingly engaged with the full (research) data lifecycle, the traditional scope of an 'OAIS' is from first point of contact with a data depositor, through data ingest, until completion of delivery of data to a user (access). For repositories providing access to sensitive data there is an increasing need to mediate the *use* of data within secure access environments (remote or safe room), and by offering research project approval processes, training and statistical disclosure review before final outputs can leave the archive-mediated access/use system.

CESSDA ERIC Archives are far from homogenous in terms of their size, governance, or technical and human infrastructure, but standard archival reference modelling and trustworthy digital repository criteria provide a basis for alignment and comparison. In the CESSDA ERIC context the members, as national archival/repository/data services, will increasingly be supported by cross-national coordination of tools and services at the ERIC level. These cross-national services will benefit from the same common standard baseline deployed by member repositories.

At the architectural level (general archival functions, activities and many standard business processes/workflows) the impact of NNfD is limited. Some archives already mediate access, whether through processes (negotiating depositor approval for projects) or through managing secure access environments (with statistical disclosure assessments and output controls). But the impact of NNfD will be felt at the infrastructure level in changes to the mix of technologies and skills required. The type and degree of impact will depend on whether the infrastructure and workflows remain predominantly 'study-driven' or adopt newer formats and technologies to offer the range of services implied by data linkage and analytics at scale (see below). The

At its simplest the archival role is to ensure continued access to clearly defined digital objects (bounded data points), of value to data users, in a persistent and citable way.

In addition to some of the Survey/Archival partnerships described below archives partner with specific research projects which permits the evaluation of new forms of data and their potential for archiving (see Digital Objects as Data Assets).

The majority of work on archiving new and novel forms of data falls into these categories at present. These can usefully be evaluated against the lifecycle and workflow perspectives here to identify whether they are scalable to full archival and associated data services.

[15] http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d3-3_research-data-centers-iso27001_guide.pdf

Survey Research Infrastructures

Nothing precludes a survey (project or ERIC) from being an information secure, trustworthy digital repository. It is simply that these notions are more embedded within archival practice and have therefore driven the identification of the key workflows.

In comparison to data archives the survey research infrastructures (European Social Survey (ESS ERIC) and the Survey of Health, Ageing and Retirement in Europe (SHARE ERIC)), and other partners represented on SERISS, (Generations and Gender Programme (GGP), European Values Study (EVS) and the WageIndicator Survey) engage with a narrower range of data sources and a wider range of lifecycle stages.

The administrative data task in SERISS those data from the following perspectives.

> **Single Access Research Projects**. Where data may be sourced from a survey, Survey RI, University, National Statistics Institute, government department, Archive or other data provider. These may work on their own local infrastructure or work with an access provision initiative.
>
> **Access Provision Initiatives.** Where efforts are made to provide the harmonized metadata, accreditation and infrastructure necessary to support national and transnational access to and use of data. Researchers apply for access to specific data which may be subject to usage restrictions (safe room, secure virtual research environment).
>
> **Environments with statistical disclosure control.** Where data contain identifiable information, or at risk of identification during the research process is subject to statistical disclosure risk analysis and an approval process before outputs can leave the controlled system.
>
> **Integration in social surveys processes.** Matching additional data to survey microdata, before or after the survey process has the potential to enrich the research data available. The ESS and GGP examples, (discussed in Deliverable 6.2.1), exemplify an ex-ante integration, while the SHARE project exemplifies an ex-post process.

The focus above on the provision of specific functions reflects the mission and goals of the Survey ERICS. This contrasts with the perspective of archives which must integrate their workflows into scalable services across a wider range of data sources for reasons of efficiency.

In data lifecycle terms the survey actors are directly involved in the conceptualise/create/capture phases: controlling the design of surveys and their implementation (usually through a third-party data collection agency). In OAIS terms the survey practitioner must deal with 'pre-Ingest' negotiations related to deposit (of survey results, and of administrative or other linked data) and the curation and secure storage of data. Like Archives, survey actors create digital objects (the survey, the survey results, the survey results enriched with administrative data etc.) for dissemination to, and analysis by, users beyond their systems. A critical difference from repositories is that for the survey data (if not for the administrative data) the survey actor has a strong influence over the details of consent provided by data subjects. Another difference is that survey practitioners can be also directly involved in the analysis of the data to derive research outputs. Various partnerships may be required to support data needs whether through high performance

computing (HPC) environments for data at scale, or with providers of secure access/statistical disclosure control systems.

Though the range of data sources engaged with by survey actors have expanded (social media, administrative data etc.) their primary goal remains to develop bespoke tool chains and processes to manage the specific data types necessary to enrich a clearly scoped survey. The equivalent archival tool chain must support a wider range of data types and formats at scale.

Interaction between survey actors and archival actors varies. ESS has a close relationship with NSD[16] for preservation and access provision. SHARE works with GESIS[17] for preservation purposes but retains access management and a parallel archiving at CentERdata in Tilburg[18], and GGP are working towards a preservation relationship with DANS[19]. The nature of survey data management means that the digital objects deposited from survey ERICs into archives are defined 'snapshots' of the more dynamic survey data. Researchers may be provided access to different digital objects from the survey directly than are made available through archival partners. In certain survey and micro-data infrastructures such as Integrated Public Use Microdata Series (IPUMS) and the Demographic and Health Surveys (DHS), access procedures are in place that generate bespoke data objects for users based on requirement parameters provided by the user; this is also a common use case for the biomedical sector.

Outsourcing and Complex Partnership Systems

The notion of *an* ERIC or *an* Archive, and their status as *a* TDR or *an* OAIS (See *Archival, Trustworthy Digital Repository Standards*) tend to imply a single, notional actor. In fact, the majority of data-related services are delivered by a range of actors: suppliers, partners and third parties. The more complex the partnership environment the harder it can be to develop, maintain and provide evidence for appropriate technical and organizational measures. These include measures to ensure that potentially valuable data assets retain the integrity, provenance and context necessary to support long-term preservation and appropriate security levels across workflows and throughout the data lifecycle. As products and services are developed at the CESSDA-ERIC level for use across member archives the CESSDA-ERIC will increasingly become a workflow actor. As the number of actors in the workflows widen, so do the challenges of ensuring a common circle of trust.

Lifecycle & Workflow Management Factors

In designing the workflows and supporting information necessary to deliver our data services there will be many different perspectives: social scientists, user interface developers, back end developers, systems architects, librarians, business analysts, accountants, administrators, data scientists, data managers, funders, curators, faculty, project managers, and archivists.

[16] http://www.nsd.uib.no/nsd/english/index.html
[17] https://www.gesis.org/en/institute/
[18] https://www.centerdata.nl/
[19] https://dans.knaw.nl/en

We need to manage technical and organisational issues that support change across a range of perspectives. Some issues will be generally applicable across a range of actors, but many will be locally specific. These factors form the local 'organisational context'.

Consent

Archives will tend to seek assurances from depositors that all data has been appropriately anonymised before deposit. Suitably anonymised data avoids a number of challenging consent issues, though may retain related legal and ethical issues if linked in a manner which might permit re-identification of data subjects. The status of data as personal/anonymised or at risk of re-identification impacts the data protection and sharing measures for associated workflows. Provenance information pre-deposit remains important to end users of the data e.g. consent for research purposes not sought at point of administrative data collection, can be sought at the point of survey participation, but the related risk of consent bias should be documented. Mechanisms and outcomes of consent behaviour are outlined in the figure below as referenced by (Künn 2015)

Figure 2. Mechanisms of consent behavior

Mechanism	*Description*	*Hypothesized outcome*	*Empirical evidence*
Uncertainty	Respondents give consent because they cannot recall the required information—they allow the requested information to be drawn from administrative records.	Consent	No
Confidentiality concerns	Respondents fear that administrative records contain sensitive personal data and might be misused.	Denial	Strong support for negative relationship
Resistance to the survey	Respondents have a general resistance to the survey (they distrust the institution behind the survey), have no interest in the survey, or otherwise are opposed to the survey.	Denial	Support for negative relationship
Relationship with the data-owning agency	Respondents have a relationship with the institution that provides the administrative records; for example, they receive benefits or services from the institution.	Ambiguous	No support for a direct relationship, but the relationship to a government institution has positive effects
Attentiveness	Respondents are impatient or inattentive and are not willing to read or understand the consent form.	Ambiguous	Support that inattentive respondents face a higher probability of giving consent
Interviewer effects	The characteristics or views of the interviewer may influence respondents' consent behavior.	Ambiguous	Little support

Source: Sakshaug, J. W., M. P. Couper, M. B. Ofstedal, and D. R. Weir. "Linking survey and administrative records: Mechanisms of consent." *Sociological Methods & Research* 41:4 (2012): 535–569 [9].

IZA World of Labor

Figure 1: Mechanisms of Consent Behaviour

Gaining informed consent for archiving and data sharing should be seen as 'one more small step' to gaining consent from participants in a research project' (Summers 2017), with the recommendation that the option is provided on a granular level to consent to "what will be included in the archiving process" and to "identify and explain the possible future uses of the data".

Digital Objects as Data Assets

To curate our data we must define the *digital objects* we are managing including any associated structured metadata and other documentation. The workflows themselves are supported by curation process metadata. NNfD differ from the 'standard' survey/aggregated statistics or qualitative data which the CESSDA organisations typically hold in a number of ways. Identifying the characteristics of transactional data and social media data, experimental economic data, crowd-sourced data, consumer data, mobile sensor data etc. supports the creation of object models that can be used to define the implications of NNfD, including the re-tooling of any technical infrastructure.

Each digital object is effectively a boundary put around some data points for some purpose. The primary 'traditional' target object addressed by SERISS is 'the Survey' which implies data, metadata and documentation (the simplest object model) which must be managed over time and through multiple connected waves of data collection. Understanding these features of the object helps us define what metadata we need to manage it through the lifecycle such as file formats (and dependent technical environments) clear time-stamping of actions, versioning, change logs and structural metadata. These allow us to relate originally deposited digital objects to the versions of those objects designed for dissemination to researchers or designed for long term digital preservation.

Linked data formats such as RDF imply an extension to the traditional ideas of clearly defined sets of files moved together from one storage location to another. Volume has an impact on selection and appraisal decisions and on the infrastructure needed to support all levels of curation (see Ingest), selecting data from fast moving data streams make defining an object, and its underlying context, harder. We're facing a far wider variety of data sources and formats which present curation and preservation challenges including quality assessment issues. These 'objects' (from a user perspective) may be associated resources held in a variety of locations with a variety of data management practices. Linked data objects imply greater reliance on structural metadata and less straightforward approaches to 'Archival Storage'.

Whether the data is stored in the *same* object as the associated metadata/documentation (a pdf data dictionary or a licence) or as the process metadata is a local decision. We are moving from a world of data objects which are primarily 'study-driven' to a 'mixed ecology' where study data sits alongside other data sources not originally intended for research (social media, administrative etc.). This mixed ecology includes the concept of placing both study-driven and new forms of data which have been processed (flattened) into a *data store (See:* Technical Infrastructure*)*. This moves us from 'discrete storage' of 'separate objects' towards the possibility of greater data linkage within a single data 'lake'. Traditionally our database models only store data that has been modelled/structured, while a data lake stores it all—structured, semi-structured, and unstructured.

Structured	Position Data	CRM	Financial Data	Loyalty Card Data	Helpdesk Tickets
Partial Structure	Email	PDF Files	Word Processing Documents	Spreadsheets	RFID Tags
Semi Structured	GPS	Web Logs	Photos	Satellite data	Social Media Data
Quasi Structured	Blogs	Forums	Click-Stream Data	Videos	XML Data
Unstructured	Mobile Data	Website Content	RSS Feeds	Audio Files	Call Centre Transcripts

Table 2: Data Types by Level of Structure

Note that for the foreseeable future archives will remain a mixed ecology of technologies, files and native linked data. Traditional data curation paradigms will be required as derived products output by curators and/or researchers may be redeposited as discrete *study* data (file) objects in the same or in other archives. Archives must address the issues around costs, skills and interoperability when migrating to and maintaining a mixed technical ecology.

Note: Digital objects, derived products and their associated metadata requirements will be expanded upon in the next iteration of this deliverable alongside more detail on appraisal and selection for a range of specific data types/sources and the challenges of versioning complex, linked objects.

Emerging evaluations of NNfD for archiving

Archives a familiar with the challenges of archiving both survey and administrative data. The surveys themselves are at the forefront of evaluating the integration of administrative data for the purposes of research. Though not designed with research in mind administrative data sources often meet the desirable criteria for archives of having a clear provenance and integrity chain with integrated quality control; often supported by clear structured metadata. But at present the research community is in the early stages of evaluating web and social media data and few of these efforts have progressed to the point of being archived for long term preservation.

Initial efforts in the evaluation of social media data have focussed on Twitter data.

“Twitter is particularly compelling because of its perceived accessibility. In comparison to Facebook, which is largely closed-off to the academic community, or a high-bandwidth site like YouTube, tweets are small in size, public by default,”

Though Twitter data is already used across presentations, television, political newspaper coverage and websites there is seldom clarity on how the data were collected, stored cleaned and analysed (Driscoll and Walker 2014). Within the academic sphere these data must be supported by additional context if they are to form part of the research lifecycle.

Some valuable steps in evaluating these data for archiving (Kinder-Kurlanda et al. 2017) noting particularly discrepancies in data collection via the Twitter streaming API, commercially available “firehose” and Search API, and a lack of common understood language to describe the features of these data; necessary as a precursor to managed metadata design. Terms and conditions limitations on archiving the tweets themselves presents a particular challenge to the traditional archive view of digital objects as identifiers

must be stored, ideally alongside a method to 'rehydrate' some version of the originally collected data from Twitter themselves.

But even this promising work remains fundamentally embedded in a file-centric paradigm with ID data exported from a MySQL Database and stored in *.zip files at the GESIS datorium[20].

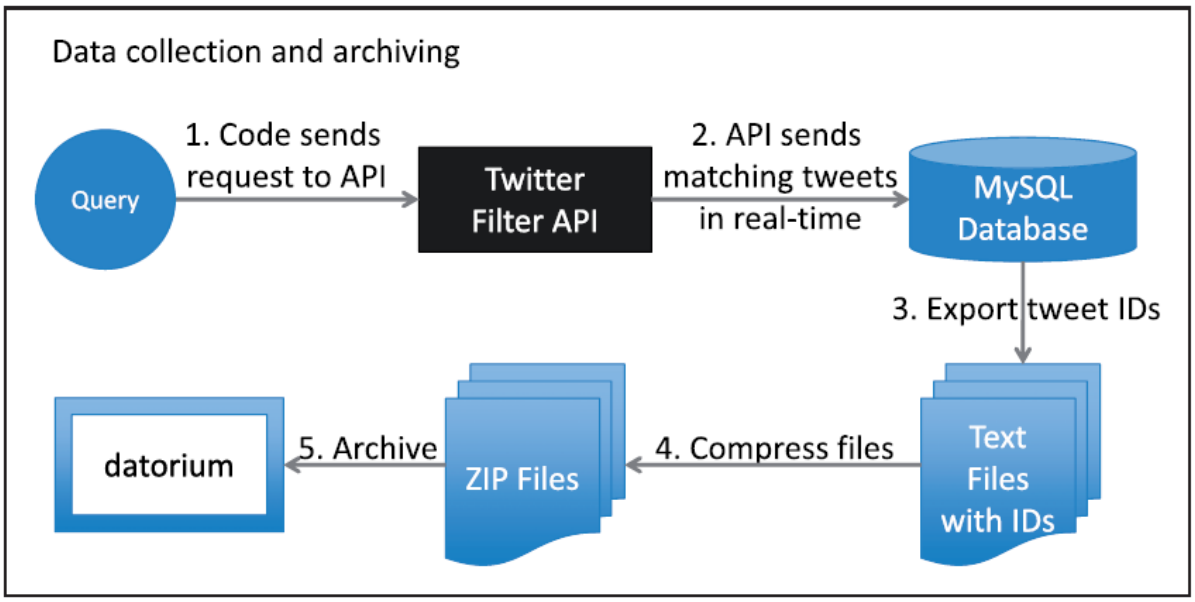

Figure 2: Schematic of the Collection Pipeline (Kinder-Kurlanda et al. 2017)

Data Linkage

Data linkage, even from familiar data sources such as survey and administrative data, has great potential for scientific and policy-related search, but the upwards trend is slowed by concerns about data security and privacy (Künn 2015). In the absence of a common unique identifier across the datasets to be linked a combination of imperfect identifiers" (names, birth information, address information etc). Data cleaning and similarity measures through probabilistic record linkage are required to prepare for linkage. A key challenge for NNfD-native environments will be to mitigate the time-consuming work needed to prepare for data linkage.

Pros

- Data linkage overcomes some of the shortcomings of the two separate data sources.
- Data linkage opens new research opportunities by combining highly reliable administrative data with detailed survey data.
- Administrative records, already collected routinely, are a cheap and authoritative source of data for enriching survey data.
- Data linkage can lower survey costs by requiring fewer questions.

[20] https://datorium.gesis.org/xmlui/

- Data linkage enables sensitive data, such as wages, to be drawn from administrative records, reducing the burden on respondents and likely lessening survey dropout and item nonresponse rates.

Cons

- Linking data can be very costly and time-consuming, mainly because of drawn-out negotiations with data providers.
- Privacy concerns and resulting legal constraints and the need for data anonymization restrict data access and content.
- Requesting consent to use the linked data may introduce consent bias (consenters differ from non-consenters) or may reduce response rates, introducing yet another selection bias.
- Sound linkage requires a unique identifier for each individual; without such identifiers, linkage becomes burdensome and matching quality may suffer.

Whether through manual/bespoke approaches of NNfD-native technologies subsets of data from multiple linked sources can be combined with additional computation and analytics to derive new "information products", which also need to be defined, described, stored, versioned and managed e.g. derived products may have their own, new levels of disclosure risk or more complex multipart rights and access requirements.

Versions for Curation and Access

Different workflows copy, amend, expand, split and merge digital objects. Appropriate versioning must be defined for a variety of situations:

- Versions submitted
- Micro-versioning to support records of curation/processing activities
- Release versions to end users
- Data products as independent objects which require versioning
- Versioning of complex collections of objects e.g. a series of longitudinal studies and its related documentation artefacts

Accurate version management is critical for persistent identification of digital objects.

Quality

Subsequent iterations of these generic high level workflows will evaluate the impact of quality assessments. Administrative records are consistently and accurately collected, resulting in highly reliable data covering a large number of observations (Künn 2015) but web/social media data may have more questionable quality levels which must be taken into account pre and post-linkage.

Different perspectives on quality must be reflected. DAMA 2013 identifies Completeness, uniqueness, timeliness, validity, accuracy and consistency as the six primary dimensions for data quality assessment, (Bruce and Hillmann 2004): completeness, accuracy, provenance, conformance to expectations, logical consistency/coherence timeliness, and accessibility..

The DAMA focus on uniqueness implies a challenge for archival workflows where multiple versions of data may be 'live' at any one time whereas the Bruce and Hillman focus on *accessibility* reflects an archival priority in line with FAIR data principles[21]

Technical Infrastructure

"The same tweet delivered by the Streaming API[22] and Gnip PowerTrack[23] will require different software to read, take up different amounts of space on the disc, and include different supplementary metadata." While the Streaming API is free of charge and Gnip PowerTrack may cost tens of thousands of dollars, neither service can feasibly be run at mass scale without significant computing resources and the cooperation and expertise of an interdisciplinary team

When the core asset is data which is dependent on a particular technical infrastructure to be selected, managed and used the curation problems are technical problems and vice versa. We need to ensure that discussions around data are conducted and understandable in business terms as well as in terms which support research and technology perspectives.

Study-driven workflows

A 'study is a collection of data and documentation files deposited with an archive, For the majority of archival collections traditional research 'studies' are deposited as a collection of data and documentation files; these are normally a single dataset with associated metadata. Standard curation processing work is undertaken on a copy of the deposited information, the study is enriched with additional metadata, including a catalogue record, followed by creation of further copies for data storage (with appropriate preservation measures) and for download access by appropriate users (in relevant formats with information which supports usability). The size and complexity of these data usually permit their direct download by users. Further research and analysis takes place outside of the archival environment.

Though data may become more heterogeneous and increasingly complex through linkage, as long as clearly bounded set of files and folders are transferred from a survey to an archive for deposit, curation and direct delivery to researchers there is little change to high-level archival workflows. For instance, deposited SPSS files continue to be forward-migrated to newer SPSS formats supported by applications on users' desktops. The challenge becomes more complex as demand rises for access to more sensitive personal data, or for more complex data analytics and linkage, which may not be legal, ethical or practical to work on in the researchers' local environment. Some archives already provide secure access/statistical disclosure checks/output approval functions for sensitive data. For less sensitive data the traditional access path remains to allow the researcher to download the files for analysis in their local environments. Both surveys and archives work extensively with traditional file formats such as SPSS while managing metadata (for curation process tracking, integrity, resource discovery etc.) through relational databases. But for NNfD the curation approaches applied by an individual survey (bespoke tool chain, manual intervention etc.) may not scale to meet the range of data deposited at an archive using SQL databases and SPSS files as their primary technologies.

[21] https://www.force11.org/group/fairgroup/fairprinciples
[22] https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data
[23] http://support.gnip.com/apis/powertrack/

Addressing New and Novel Forms of Data

Though the underlying architecture of archives remains familiar, the NNfD, not previously available to research, have implications for infrastructure including a requirement for improved data manipulation (*analytics, linkage, visualisation, statistical output/disclosure evaluation*) and increased process automation to handle a wider variety of data at scale. Mediating access to these data implies user-facing environments beyond presentation on web pages or access through downloads (*Secure Access, Safe Room, DSaaP, etc.*).

We need to support nonlinear data structures, when data elements are not organised in a sequential fashion. Data structures like multidimensional arrays, trees and graphs are some examples of widely used nonlinear data structures and these are a common feature of new and novel forms of data. For example graph structures are often used to represent connections in social networking sites and tree structures are used to represent hierarchical data relations such postal addresses, IP domain names and genealogies.

Increased frequency of deposit, of large linked-data derived from a greater variety of formats present a serious scaling challenge to archives with a file-based study-driven approach. Indeed, the notion of a "file" may become anachronistic when content is fragmented and distributed across a thousand machines in a cluster and replicated several times for redundancy. The kinds of 'big'/NNfD linkage and analytics increasingly in demand from data users (including survey actors and survey data users) are the same sort of analytics necessary to support the curation workflows for quality assurance, data normalisation for linkage, linkage risk analysis and statistical disclosure control.

The technologies which support the scaling of traditional study driven workflows to 'big'/NNfD and which support the delivery of new and novel forms of research promised by data analysis and linkage at scale are increasingly available and accessible to the research community. In recent years they have reduced in cost and in the level of specialist expertise required to deploy them. At the UK Data Archive the Data Services as a Platform (DSaaP) initiative has adopted the ODPi (Open Data Platform Initiative)[24] as a reference platform to support enterprise-class big data solutions, based on Hadoop and other big data workflow technologies such as Apache NiFi[25]. ODPi-based technology has a number of defining characteristics which radically simplify the way data is stored and accessed to enable both data and its associated metadata to be linked together, and to describe the types of relationship that exist between and within them. The conversion of deposited data to neutral formats commonly accessible to these system (e.g. RDF) permits us to develop workflows which map to traditional archival architecture but which are deployed on new infrastructure to support the scaling of deposit and curation. This approach also provides for data sources in RDF which natively support rich data linkage and the generation of derived data products. These data products may be developed during archival curation to streamline end user research, or may be generated by the researchers themselves.

Workflows and our ability to change and scale them to handle NNfD will be directly impacted by the mix of technologies available to the archive, whether on premises, cloud-based or through some partnership model.

[24] https://www.odpi.org/
[25] https://hortonworks.com/products/data-platforms/hdf/

Preservation/Long Term Access

The scale of big data, and the opportunities for generating derived, linked data products offered by NNfD technologies, my mean that not all data and data products can be stored for the long term, or even reproduced from past 'queries'.

For example, survey response data related to fuel poverty and energy consumption data (both potentially sensitive), may be linked to geographic data on elevation or local temperature. If undertaken at scale the resultant data products may not all be amenable to long-term storage. If one data source is streaming and not retained for the long term it may not be possible to re-run the linkage at a later time to reproduce the data product.

Therefore the opportunities derived from these new data platforms are accompanied by challenges about which data sources and derived data products can be retained/recreated and how these more complex objects might be versioned, persistently identified and cited from resultant research. The preferred archival preservation default position of 'keep everything' will need to accommodate new trade-offs driven by the practicalities of sampling resolution. Active decisions about which data products are to be retained for the long term must be taken and the nature of that long term availability must be transparent to data depositors and users. In the interim, a cautious approach to retention of a range of data products may be advised while long term value and preservation needs are assessed. There is a (reducing) cost to data retention, the cost of inaction may be higher.

People, Roles and Skills

New legislation requires new processes ('measures' in GDPR terms) in our technical and organisational environments. We need to define the roles necessary to support these and the associated skills required. The technical skills and associated roles for workflows need to develop as we address NNfD. The need to scale will also transfer some activities from manual, to human mediated to automated processes as the relevant technologies are deployed.

Legal and Ethical Context and the GDPR

The GDPR introduces Data Protection by design and default, andencourages establishment of data protection mechanisms and of data protection seals/marks for the purpose of demonstrating compliance with the GDPR ("codes of conduct and certification", Article 40-43). All actors across the workflows must define technical and organizational measures to provide appropriate protection to data, including the assignment of actors with specific roles in workflows, such as data controllers . Even in cases where activities are technically lawful there remains the possibility that ethical considerations will preclude them.

Changes to traditional technical infrastructure and new infrastructure (e.g. data lakes) will require Data Protection Impact Assessment (DPIA) (Article 35).

In the vast majority of cases data deposited with archives in anonymized, but both archive and survey actors need to be confident that any additional processing actions they undertake or facilitate are either within the original bounds of consent, are permitted under another legal basis (Article 6, 2 (e) or Article 9, 2 (j) may be used when carrying out a task of public interest) or permitted through an appropriate derogation. Processing of personal data for archiving purposes in the public interest/ research/statistical purposes shall in accordance with Article 89 (1) (appropriate safeguards to ensure rights and freedoms of data subjects) not be considered to be incompatible with initial purpose (Art. 5 (b). However,

Article 6, 3 states that public interest as lawful basis shall be laid down by Union or member state law hence the conditions for using “public interest” may vary between countries. Personal data may be stored for longer periods for the same reasons as above (art. 5 (e)).

Unless we address these issues to maximize the potential for re-use our ability to derive additional value and impact from the data will be reduced. There is also an implied need to develop extensive data subject contact management processes to notify of new processing, handle withdrawal of consent, and seek additional consent. Requests for data deletion may reduce the value of historical data and preclude verification and replication of past research. The nature of a derogation does not depend on whether an actor is a ‘research’ or an ‘archival’ organisation, it depends on whether the activity (part of the workflow) undertaken is for an archiving or a research purpose. Survey actors undertake processing in preparation for archiving, and Archive actors undertake processing for the purposes of providing access to research.

There are four distinct bodies of literature which, in combination, are used to explore the legal and ethics of research using big data and/or social media data:

- Legal texts/ requirements, including Terms of Service (ToS) of social media providers;
- Recommendations about ethics from supervisory authorities, at EU level and Member State, for example, the European Data Protection Supervisor’s Code of Digital Ethics;
- Professional and practitioner guidance on ethics of social media research, such as the Codes of the Association of Internet Researchers, of the British Psychological Society (academic) and ESOMAR, or IPSOS-Mori, or DEMOS (practitioner). Within this corpus, the standard Codes of Research and Professional Ethics are also considered;
- Academic literature on research ethics of big data and social media data from a range of disciplines (sociology, psychology, health), broadly conceived as ‘internet research ethics’.

Data Lifecycle Perspective

The data lifecycle functions presented here and under Generic High Level Workflows below are based on the DASISH data/metadata quality lifecycle. The more detailed workflow items have been derived from work within the UK Data Archive teams to compare and contrast their traditional workflows with those implied by the NNfD-native DSaaP product. The functions and workflows presented are those required for the UK Data Archive and UK Data Service to develop and manage a hybrid technical infrastructure.

> *“Researchers increasingly want access to detailed information about what happens in the early phases of survey design and implementation: the sampling process, questionnaire development, fieldwork, post-survey data processing, weighting – basically all elements of the survey data life cycle. It is clear that knowing more about what occurs along the life course of a survey sheds light on the quality of the resulting dataset.”.*

Though the primary focus for workflows is the archival functions (deposit to access/use) it follows from the above that workflows for NNfD must address this demand for full lifecycle

information. Our selected approach has been to adopt, and adapt as necessary the metadata quality lifecycle model developed for the DASISH project.

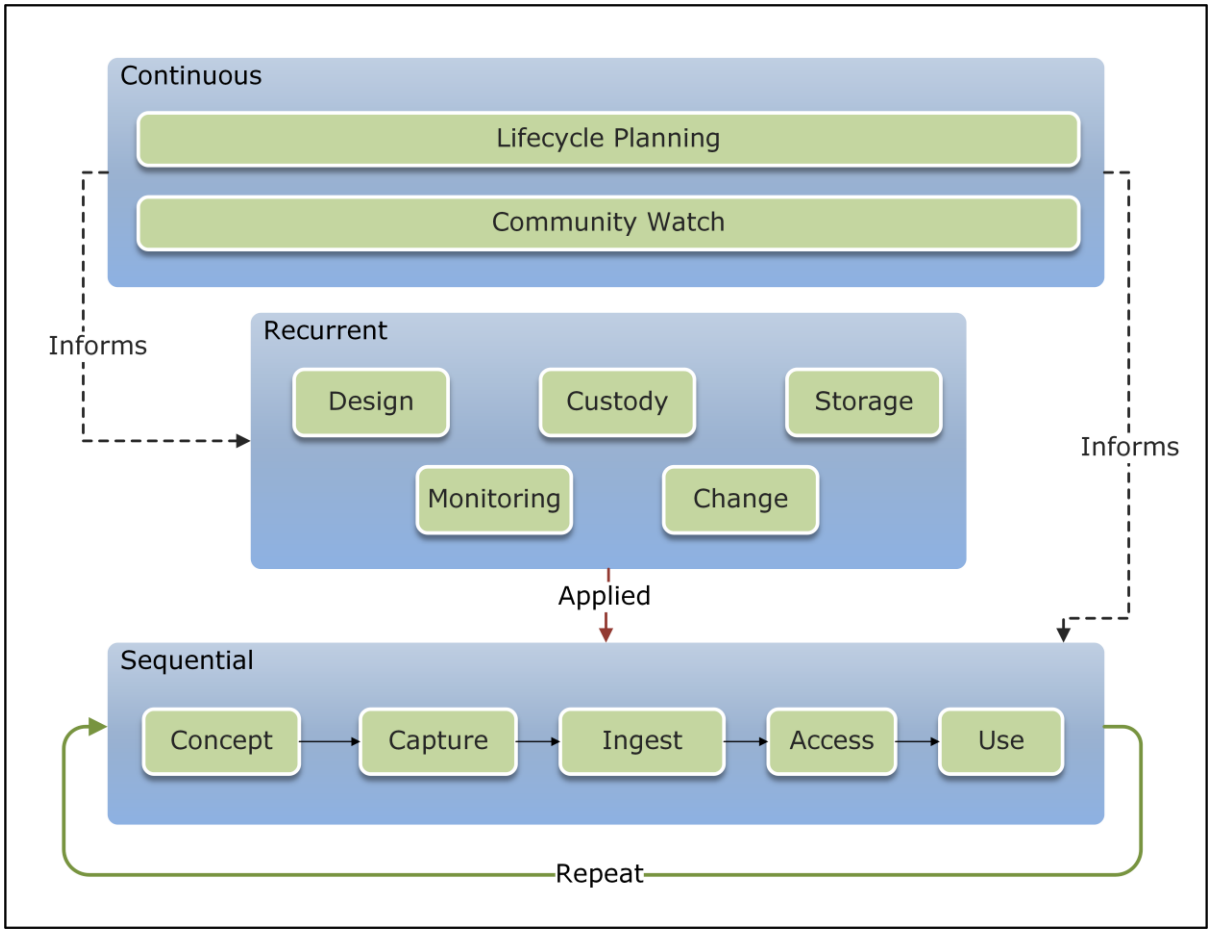

Figure 3: Functional Lifecycle Model.

Together the elements of the functional lifecycle provides an architectural blueprint against which existing or planned infrastructures can be mapped. Each actor in the lifecycle should understand the functions they take some or all responsibility for.

Continuous

Processes which continue throughout the digital object lifecycle.

Lifecycle Planning

Planning for curation across the whole digital object lifecycle.

Each actor should understand their role and responsibilities in each lifecycle phase and what they can expect from other actors. For complex partnership environments necessary to support NNfD this presents data governance and cooperation challenges (cf: circle of trust above)

Community Watch

The OAIS model references the need to develop technology watch activities, with particular regard to ensuring that data is stored in formats which have a low risk for long term preservation and accessed in formats that meets the need of the users. But in addition to technological advances (and obsolescence) there is a need for a wider process of

monitoring and responding to a range of factors including stakeholders, product vendors, standards bodies and legislation. Organisations like the CESSDA-ERIC are in a position to coordinate such activities across members, though there will always be issues that are entirely national or local.

Community Watch might include interactions with stakeholders (researchers, funding bodies, potential depositors) during the conceptualise and create/capture phases of the lifecycle (see below), i.e. prior to the beginning of a formal negotiation for deposit (see Pre-Ingest)

Many repositories undertake research data management training of researchers, provide guidance on deposit and are involved in wide-ranging stakeholder engagement (through training, events etc.) with potential depositors. These activities can be vital in finding new sources of data and in maximising the quality of data offered for deposit. Good quality curation by the original data creator is always preferable to the same curation undertaken by the archive as the originator has the most direct knowledge of the data. This provides for improved data being made available to users and reduces the cost of curation to archives.

Recurrent

Functions which benefit from being designed and considered centrally, then deployed consistently throughout the lifecycle.

Design

How the infrastructure is designed and documented, and how it is redesigned in response to changes in the stakeholder ecosystem or organisation context.

The design process informs some of the community watch activities and responds to identified changes.

For example, in the case of adopting Open Source technology to support NNfD this may involve adapting to frequent release cycles, but also issues with forking, adopting or supporting a non-mainstream fork, or mitigating the risks of documentation not being aligned with the software release cycle.

The workflows themselves provide the initial design state and are artefacts for change management.

Custody

Custody and custody transfer, including deposit, ingest, access, and potentially internal custody transfers.

It must be clear which actor has responsibility for the data at each point in time. Custody transfers are critical points in workflows where data integrity or provenance trails are at risk.

For NNfD products like Apache NiFi support greater machine actionability and process orchestration. Movement of and changes to data sets are recorded at data point level and auditable in a way that could be prohibitively resource-intensive on Windows file-based systems. Volume, in this sense, is not simply a feature of big data, it is a feature of big data management as well.

The workflows consider the point of custody transfer from depositor to archive and from archive to user. Archives with complex internal staff arrangements (multi-partner/multi-site)

and/or third party services may need to define and manage additional points of custody transfer during the workflows.

Storage

The focus on lifecycles and workflows should not detract from the practical need to define a clear data pipeline. Each archive must clearly define the data deposit and access/use routes and the storage locations which underpin each workflow. Each element of the data pipeline must have integrity, backup/restore and information security measures appropriate for the data and workflows involved. Backup procedures may need refactoring to take account of a less file-centric storage infrastructure. For example, “backing up” 5 petabytes of data onto tape may simply not be practicable and new models of data “availability” based on concepts of hot, warm and cold storage across distributed environments with different cost models will become more prevalent.

Documented data storage, including the management of integrity-checked objects before and after any change is required across the workflows. If we have long term preservation responsibility or hope to provide access beyond the next round of change, we also need to have archival storage that is responsive to changes in technology and formats.

NNfD may affect this function. Rather than just outputting objects for access and use the system may need to support the integration of derived data objects either as new objects or as repeatable queries (so the object may be re-derived in future), as well as identifying constraints on reproducibility (e.g. if the data against which the query was run is not retained). In a multi-petabyte world, consideration will need to be given to whether it is meaningful or possible to store every copy and version of the data during its lifecycle.

Monitoring

Continuous monitoring to support appraisal and disposition across the digital assets and processes.

Appropriate technical and organisational measures must not only be developed, but also monitored over time. Effective collection of business information is necessary to support a range of internal and external assessments, (quality, risk, maturity), and analysis and reporting.

Archives which are under audit e.g. for TDR status or information security, must generate appropriate artefacts to act as evidence. Some evidence may relate to the day to day data processing workflows processes: deposit licences, user licences, provenance change logs, quality test outcomes (pass/fail of virus or integrity checks for instance).

NNfD-native environments may offer data and activity monitoring by default, avoiding some of the risks of human errors associated with manual data processing.

Change

Change and change management in this context does not relate to amending and versioning digital objects but to moving from one organisation context to another. The redesign of a workflow based on findings from community watch must identify which change management mechanisms are required to ensure data are not put at risk.

Archives which are under audit e.g. for TDR status or information security, must generate appropriate artefacts to act as evidence of good change management practice including Data Protection Impact Assessment (DPIA).

Sequential

Sequential and repeatable functions,

Concept

The sequential lifecycle phases begin with the first initiation of an idea about the data to be created or captured to some purpose. In traditional study driven research there are numerous benefits to capturing funding metadata and recording the study-design process from an early stage.

For NNfD the data of interest to researchers may have been developed without research in mind and there may be very limited context about the integrity and provenance of the data prior to capture.

Capture

Capture and/or creation of data assets.

For the study-driven workflow archives are seldom involved in the create/capture process and must seek to gain as much context as possible from depositors.

For NNfD the researcher may be undertaking exploratory data analysis, creating linked derived data products in a system under the control of the archive with full provenance tracking. But as noted above the more likely case is that data are collected or generated without the benefit of researcher design expertise or a clear data management plan.

The remainder of the sequential lifecycle functions are covered below as part of the traditional archival phase of the data lifecycle.

Generic High Level Workflows (for Archives)

The GDPR requires data controllers and processors provide clarity and transparency to data subjects about how and why their personal data is being processed (Summers 2017). Clear descriptions of data workflows and processing will support such transparency and the justification of any particular derogation.

These high-level generic workflows for new and novel forms of data are defined as generally applicable groups of activities which apply to archives, but also to other data stewards against which the impact of NNfD may be considered. The high level workflows also provide a reference point for designing change management to move between the current and the future data management paradigms.

The agreement of high-level, stable components is important due to the limited examples of NNfD being archived at scale. Existing survey infrastructures work on goal-focussed linkage with (primarily) population data at a (primarily) human-mediated level. The current work on Twitter and other social media data is informative and valuable in terms of identifying data challenges (limited data provenance, integrity risk etc.) but also remains primarily a bespoke process conducted by a small number of subject experts at (relatively) small scale. The current NNfD work within survey infrastructures and among social media researchers can

indicate specific data quality challenges but cannot yet define the impact on archives of curating high volume, high velocity, high variety of data at an operational level. The final reporting phase of the Big Data Support Network and the development of their Data Service as a Platform (http://dsaap.info/) has provided a number of insights. One implication is a shift in the balance between manual processes and machine-actionable processes and the required commitment to develop ingest, access and statistical disclosure protection models that can be operationalised by these NNfD-native platforms e.g. privacy engineering.

The functional headings below may be characterised as 'internal' (e.g. curation) or 'user facing' (e.g. access). All internal functions contribute to user facing functions in some way, but have different audiences and communications requirements. All functions require that their stakeholders and actors, including data suppliers and consumers are clearly defined and managed.

The research data community is increasingly accessing 'new and novel forms of data' not originally intended for research including social media and administrative data. While the archival functions (i.e. the architecture) remain fundamentally the same, the challenges, including those around personal and sensitive personal data, are significant. In some ways, the data themselves do not present unique challenges as repositories have always had to face increases in scale, complexity, etc.; it is the new possibilities for a new generation of data scientists presented by these data, and the accompanying demand for new and novel research technologies, that pose challenges to the existing archival infrastructure.

NNfD have a wider variety of sources and their custody history, provenance, quality, integrity etc. may all be harder to define. For example, data linkage at scale and links with data from other domains such as the biomedical sciences, will require new approaches and the development of machine-actionable mechanisms to partition sensitive and non-sensitive data at a much greater level of granularity than has traditionally been the case.

Our approach necessarily includes the assumption that some data will contain sensitive personal information. Once the archive provides a safe room or secure remote access facility is mediating the use of the data (research) rather than mediating access (archiving). In terms of the GDPR framework it will be important to identify the areas of activity where research rather than archiving derogations apply. Workflows that act on personal data will also need to incorporate the possibility of withdrawal of consent and notifications to data subjects in the case of compatible additional processing.

The full, high-level, sequential, functional view covers data's journey through: concept, capture, ingest, access and use. The subsections below start to break out the traditional archival functions (ingest, access and increasingly, support during the use phase) into more detailed, but still generic, functions and areas of activity which can be applied across a range of repositories, archives and other trusted data stewards.

Pre-Ingest

Activities from the first point of contact with a potential depositor, through negotiation, deposit and access licence agreement, quality assurance etc., until the deposit is handed over for ingest curation.

Pre-Ingest and Ingest undertake the transition of data assets, and custody of those assets into a trusted environment.

Pre-Ingest covers the acquisition of data for the archive. It includes:

- the management of contacts as people and organisations with roles in processes, and correspondence and negotiation with those people and organisations
- the definition of deposit licence terms including the rights and obligations the archive undertakes for curation, preservation and access
- the definition of access licence terms appropriate to the type and sensitivity of the data
- the agreement and validation of acceptance criteria for the particular data to be deposited
- Assessment of disclosure risk of the data and any mitigation measures to be performed prior to deposit

An acquisition may be an entirely new deposit, or related to an earlier deposit (e.g. a change/correction or an addition). Whether we treat new data as new objects or as changes to existing objects depends on our object model; this is how we define the boundary around data, and any grouping between objects (e.g. grouping longitudinal waves, or grouping by topic). Unless there is a compelling justification data and metadata are not overwritten, so we need a versioning approach if we change our objects over time.

In an ideal world a deposit is a single, clearly bounded set of data points with associated documentation and metadata. In reality the deposit may arrive in chunks, over a long period, or out of sequence, and need to be incorporated into an ongoing acquisition 'object'. For NNfD with potentially complex deposits at scale each archive needs to communicate with depositors to define transfer events, manifests of data transferred and clearly agreed points where custody is transferred.

Given the risks of accepting custody of data of uncertain provenance and the cost of ingest the pre-ingest process is likely to take on an even more critical role for NNfD. For highly complex data sets which may be linked to others within the archival boundary (by curators or users) it may be challenging to identify what data can be retained and for how long. The level of preservation to be undertaken must be clearly communicated to the depositor.

For NNfD pre-Ingest workflows may work with a wider variety of deposit pathways including some which are purely-machine mediated. For example, the continuous deposit of smart meter data every few minutes will still require a contractual agreement and deposit schedule. Streams of data must be continuously validated, or at least periodically sampled, to ensure they continue to align with the deposit agreement to avoid, for example, accidental inclusion of personal data in a stream which should be anonymised. In such cases, the traditional notion of a "depositor" may need to be re-defined. For example, streaming device data will originate from potentially millions of households, each householder providing the consent to use that data. Although there may be a single point of ingest into the archive managed by a national infrastructure body, there is no single actor who fulfils the role of "depositor" as understood in a traditional study-driven context.

Data streams challenge the traditional technical infrastructure view of file-based digital objects.

Depositor Account

The availability of a depositor account user interface changes the balance of which processes are managed entirely through back office staff corresponding directly with depositors, and which processes can be rapidly handed over to the depositor for decision/action. This impacts efficiency, cost, and accountability during pre-ingest workflows.

Depositor Contact Management Workflow

Contact management for handling people and organisation metadata related to data creators and depositors. This may begin before an acquisition has been started.

Applicants for the CoreTrustSeal will often note their close relationships with depositors over time. The extended range of sources for NNfD will change the nature of these relationships and necessitate more rigorous depositor identification.

Depositor Correspondence management workflow

Managing the contact trail between an archive and a contact both generally and in the context of particular acquisitions. This may begin before an acquisition has been started.

In GDPR terms the contact information and correspondence may contain personal information whose retention over time may need to be justified. Depositor information necessary for service provision (signed licences) will form part of the archival information, other contact/correspondence information may need to be periodically deleted in line with a corporate records retention schedule.

Deposit Negotiation Management (Appraisal & Selection) Workflow

Once an acquisition has been started the deposit negotiation management processes provide archival provenance through contact, correspondence, deposit and access licence criteria, and deposit validation.

Information gained and decisions taken during this process affect subsequent workflows and data flows. E.g. whether the approach is self-deposit or curated, or whether the data sensitivity requires particular storage and access methods.

Archives traditionally undertake an appraisal and selection process to ensure the data being offered align with agreed collections development principles. This contrasts with a research project or survey infrastructure which is identifying a particular data source for a clear, immediate purpose. The archive must evaluate the potential value of the data in terms as compared to the cost of curation.

For NNfD the increased variety of data sources and a lack of provenance will directly affect the appraisal and selection process. Archives may need to be involved in a wider range of relationships beyond the traditional academic sphere and investigate data sources further to ascertain their level of value and quality. The greater the resource needed to ingest process the data the more important that the appraisal and selection process is robust enough to minimise risk in the workflows that follow.

For traditional file-based research study deposits the appraisal of traditional data sources from familiar depositors may be a matter of opening SPSS files on local desktop. For NNfD there may need to be additional contractual assurances and automated processes before accepting custody.

Pre-Ingest evaluations will include:

Data quality - raw data vs. cleaned data and measurement error (identification of the percentage of errors which can be cleaned by automatic tools); Quality reports (especially with regards to Official statistics) – sampling, coverage, comparability (of different sources, across time and geography), theoretical concepts and measurements; time dimension (frequency, longitudinal character), units (minimal geographic unit, networks, individuals, other). For NNfD at scale these features must be presented through some machine-actionable metadata standard rather than in prose documentation. This will present a challenge to workflows which currently require minimal structured descriptive metadata (Dublin Core, DataCite etc.).

Potential for reuse- NNfD are increasingly used, and linked, in a multidisciplinary context where designated community and disciplinary boundaries are blurred – medical vs. social science use; humanities (e.g. history, linguistics) vs. social science use of twitter data. Local expertise must keep pace with data sourced from new domains and disciplines if the resultant data are to be usable by the designated community.

The most critical pre-Ingest evaluation in workflow terms is whether the data contains personal information. This may impact accept/reject decisions based on risk, and will define the appropriate information security protocols and the required licences needed to support access and use workflows.

Depositor Licence Management Workflow

The definition of deposit licence terms including the permissions, prohibitions and duties the archive has during curation, preservation, storage and access. This covers both any prose licence text, and any licence metadata. The latter may be machine actionable. For NNfD workflows which depend on multiple systems handled by multiple partners, machine-actionable metadata may support assurances that the data remain within a particular geographic region or that sensitive data is stored and curated in appropriately information-secure environments.

Access Licence Management Workflow

The definition of access licence terms appropriate to the type and sensitivity of the data, including the permissions, prohibitions and duties agreed by the end user of the data. This covers both any prose licence text, and any licence metadata. The latter may be machine actionable. For NNfD the provision of prose licencing may be insufficient. A clear, machine actionable access model may support automated disclosure risk analysis of data linkage and drive the application of access requirements for derived data products.

Deposit Validation Workflow

The agreement and validation of acceptance criteria for the particular data and metadata to be deposited. May include virus checking, a file manifest, checksums, format validation (against approved list) and other quality assurance activities. All of these may have infrastructure (human, workflow technical) implications when presented with NNfD.

Ingest

Ingest begins when deposit negotiation is complete, and custody for the data is handed over from the depositor to the archive and ends when the curated data/metadata is handed over to other archival workflows.

In OAIS terms: from the point a deposit becomes a 'submission' information package (SIP) to the point of storing 'archival' information packages (AIP for long term digital preservation) and 'dissemination' information packages (DIP for access)

NNfD may have familiar curation levels. The definitions below are taken from the CoreTrustSeal requirements[26]:

- As deposited- content distributed as deposited
- Basic curation – e.g. brief checking, addition of basic metadata or documentation
- Enhanced curation - e.g. creation of new formats, enhancement of documentation
- Data-Level Curation as C above but with additional editing of deposited data for accuracy

For NNfD it is clear that levels C and D may not be possible without dedicated technical systems cable of analyses at scale. For NNfD supporting systems which 'flatten' data into commonly addressable formats like RDF for linkage there will always be a curation at level C and above.

The 'V's of Big Data provide a starting point for thinking about what the new and novel challenges are and what the impact on curation will be.

Volume:

Repositories have always handled changes in data scale which impact our ability to curate and which necessitate changes to infrastructure. But the unprecedented rise in volume of NNfD may be more challenging to Ingest and require different technical skillsets.

Velocity:

Moving from a single study deposits to multiple longitudinal deposits are velocity changes, but NNfD may present velocity at a new and challenging scale, where the speed and periodicity of deposits precludes curating data manually. Handling streaming data is expected to be one aspect of NNfD that will present some challenges.

Variety:

NNfD may mean we need to go beyond a carefully approved and managed range of acceptable data formats/file types/schema. Indeed, the core NoSQL paradigm of "schema-on-read" is a profound challenge to traditional file-focussed data management.

Veracity:

Repositories have always sought to maximise their knowledge of data provenance, including producer/depositor identification and evaluation of data. Identifying quality issues across a range of large scale, new and novel forms of fast moving data from new sources presents new challenges.

Value:

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

Value would ideally be identified at pre-ingest, but the other V's might make it harder to define what the value and impact of the data are. Enabling more non-hypothesis-driven exploration of data products for users' means that identifying value may increasingly be an emergent property of use, as well as an evaluation by the curators.

All of these factors, and the increased legal and ethical challenges of handling sensitive personal data (data minimisation, sub-setting, quality assurance, risk) may affect ingest processes. In the case where current processes (human or machine) and skills/capabilities cannot be scaled to meet the challenges we must either provide new machine-mediated solutions, or communicate to users that previous levels of service (data quality, delivery frequency, delivery speed, etc.) cannot be met.

Data Curation

An Ingest workflow is comprised of predefined actions, and/or of custom actions undertaken on data and metadata, to maintain the archival provenance chain during ingest. This may include quality assurance, annotation/correction, data enrichment or data transformation. All of the workflows below may need to be amended to deal with NNfD. One of the most basic tenets of ingest is that work is undertaken on a copy of the original data, but for NNfD at scale the local environment may not support this and the detailed tracking of any changes to the data becomes more critical. In this case workflow actions which are not reversible must be identified and analysed from an integrity risk perspective.

Validation Workflow

Validation of the data and metadata deposited against pre-defined criteria. These processes are traditionally more rigorous and detailed at the Ingest stage than the deposit validation undertaken during pre-Ingest but there may be an argument to move some additional validation of NNfD (whether whole deposits, or samples) to Pre-Ingest if this avoids a significant cost to moving inappropriate data into the Ingest workflows.

Annotation and Correction Workflow

Depending on the results of quality assurance we may annotate data or create corrected versions. The approach will depend on the level of curation agreed and/or ongoing contact with the depositor.

Enrichment Workflow

Without touching the deposited data and metadata we may add context, semantics, provide background information, link to related resources etc.

Transformation Workflow

Changing the structure and format of the data to meet some purpose such as interoperability, accessibility or long-term preservation.

Big Data Note: Ideally working on a copy of the deposited data, but may be precluded by volume.

Data/Metadata Publishing

The ingest processes ends with the delivery of data/metadata to access (i.e. data publishing) and storage (e.g. archival storage) locations (in the OAIS access received versions to be published from archival storage). A subset of the metadata is also made available to support *access management* by the archive and *resource discovery* by the users. Within a

traditional, single-site archive this handover may be a simple workflow transition, but if the ingest, storage, resource discovery and access are handled by multiple entities in multiple locations more formal custody transfers may be necessary.

Text Box: Negotiating Access and Mediating Use

The workflows below address data access and data use. The most notable trend away from the traditional (OAIS) archival architecture is the increased mediation of data discovery, access and use within an archive-controlled environment. This is not unique to NNfD as archives already mediate the use of personal or sensitive personal data through remote secure access systems or safe rooms with workflows defined to approve the outputs from these environments.

Provision of data linkage service provides users with the option to create 'derived data products'; these new objects may have new disclosure risk implications. Statistical disclosure methods must be applied before data subsets/new data products can leave secure environments. In this way repositories are already sharing responsibilities for managing risks which were previously with the user alone.

The traditional model of a bounded secure exploration environment (either physically in a safe room or virtually in a remote machine) is already being eroded by federated query infrastructures that enable linkage of datasets in remote locations external to the archive cf. Datashield[27] in the biomedical arena.

The DSaaP model provides opportunities for supporting these functions for a wider range of NNfD at scale. The analytics, visualisation and exploratory linkage (to evaluate risk) necessary to quality assure and curate these data are the same functionality end users will demand to support research. These users will generate derived data products that will need to be archived in turn. We would expect this consolidation of the access and use functions to support greater reproducibility and validation of research. Over time it may not be practical to maintain separate workflows for Access and Use for all data if NNfD systems are effectively mediating granular permissions 'on the fly'.

Access

Access generally assumes that the data are potentially available and that the user has discovered them via their metadata (but see Data Brokerage below). NNfD challenge the traditional archival 'request access and download data' archive approach. Not only because the data may be too large for download but also because the tools for their analysis may not be available locally to the user so NNfD will require a move beyond a 'download' or 'output to web page' access approach for data access. Addressing digital objects at a more granular level (question, variable, etc.) may also necessitate more granular access criteria. In addition, data products derived from linking other data sources may result in new disclosure risks which require (ideally automated) flagging for the application of new derived access criteria.

Some data/metadata may be accessible to all human and machine agents without any restrictions though even this lack of restriction must be declared explicitly as a permission in any licence and access model. Even requiring a minimal 'click through' tick box agreement

[27] https://www.datashield.ac.uk/

before access must form part of the access model, especially if that click through restricts machine access (i.e. in the absence of a machine-actionable data request API).

Machine mediated access may be limited to prevent a server going down (unintentional DoS) and such restrictions may need to form part of a wider access model for NNfD.

For access workflows the balance of responsibility between end-users and back office staff and systems is influenced by whether the access request is:

User Mediated: user can agree to conditions and access without archive intervention

Repository Mediated: the user must make an access request which may be approved or rejected based on criteria related to the user or the data.

Unmediated: the user can access the data freely without restrictions and without archive intervention

For NNfD containing personal information an archive-mediated approach is likely to be required to ensure that any intended research use falls within the boundaries of the original consent.

User Account

The availability of a user account user interface changes the balance of which processes are managed entirely through back office staff corresponding directly with users, and which processes can be rapidly handed over to the user for action. This impacts efficiency, cost, and accountability during access requests.

Registration Workflow

Identity management implies one or more approaches to user registration. In a distributed archive model, or in cases where users are international, some federated access solution may be critical.

Authentication Authorisation Workflow

Post-registration the system must support user identification and the application of appropriate authorisation rules to that identity, before granting access to resources and data.

Access Request Workflow

Users ability to select resources they have discovered, identify access restrictions (and whether they can meet them) and to meet those conditions for access through their own user mediation or through archive mediation.

In an NNfD system which supports exploratory data linkage there will need to be comprehensive logic and rulesets that generate automated access triggers. E.g. a user may have the privileges to access two different data sets separately but a request to link the two may trigger a disclosure risk that changes access parameters and triggers the need for an additional access request.

Access Mediation Workflow

Back office functionality to handle access requests. These may range from ensuring the user is who they claim, to ensuring they have received statutory training, to seeking the depositors' permission to ensure the intended research use is acceptable and within the

bounds of consent. In this latter case the original depositor (or designated rights holder) becomes a critical actor in the workflow.

Data Delivery/Mediating Use Workflows

The initial access process is complete when a copy of the data is in the custody of the user (e.g. download) or when conditions for mediating use (e.g. secure access) have been set up.

Mediating Use

Demand for NNfD may require a move beyond a traditional approach where users can use data locally on their own systems after download or via the web. If archives provide an environment for data analysis then they are mediating use.

Archives already meet user demand for more sensitive data through provision of safe room and secure remote access technologies. Remote access to a wider variety of data may become more acceptable to users if the user experience and availability of embedded software and tools can be improved. In this case archives will increasingly distribute access to data, rather than distributing the data themselves.

For data with disclosure risk (whether as a result of linkage or not), users will need to make output requests to take the data out of the archive-controlled environment.

Within native NNfD environments which offer sub-setting, linkage, statistical disclosure mitigation, analysis and visualisation tools there is necessarily a greater level of archive responsibility for how the data is 'used' than for data downloaded under licence.

Output Request Workflow

If the archive is mediating use of data with a disclosure risk (whether as a result of linkage or not) some user-facing system to support output requests will be required.

Output Mediation Workflow

If the archive is mediating use some activity to evaluate output requests, handle statistical disclosure risk evaluations and approve/reject outputs requests will be required. For systems which handle NNfD natively (i.e. data lake not file-driven) there will be a need to develop automated risk analyses at scale to support human-mediated output checking.

Use-Reuse Support

Archives generally identify demand for data, including new and novel forms of data, or new approaches to research from their designated community of users. Support and training during the use and re-use phase are excellent sources of information about community of current data and tools, but not necessarily about NNfD data or research opportunities.

The new tools and exploratory data analysis opportunities offered by native NNfD systems must be clearly explained to potential users through communications and training activities. These new data sources, technologies and skills also imply a need for a review of the data and technological literacy of users in terms of legal and ethical issues. By identifying and participating in these new support and training activities the archive drives adoption, mitigates risk and demonstrates the impact of these services. The archives are in a central position as new (and future) legislation, new trends in research, and political and social changes, increase the level of contact between researchers, data controllers, processors and support functions.

References

Bruce, Thomas R, and Diane I Hillmann. 2004. “The Continuum of Metadata Quality: Defining, Expressing, Exploiting.” ALA Editions. http://hdl.handle.net/1813/7895.

Driscoll, Kevin, and Shawn Walker. 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.

Georgakopoulos, Diimitrios, Mark Hornick, and Amit Sheth. 1995. “An Overview of Workflow Management: From Process Modeling to Workflow Automation Infrastructure.” *Distributed and Parallel Databases* 3 (2): 119–53. https://doi.org/10.1007/BF01277643.

Kinder-Kurlanda, Katharina, Katrin Weller, Wolfgang Zenk-Möltgen, Jürgen Pfeffer, and Fred Morstatter. 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.

Künn, Steffen. 2015. “The Challenges of Linking Survey and Administrative Data.” *IZA World of Labor*. https://doi.org/10.15185/izawol.214.

OECD. 2016. “Research Ethics and New Forms of Data for Social and Economic Research,” November. https://doi.org/10.1787/5jln7vnpxs32-en.

“OECD Expert Group for International Collaboration on Microdata Access- Final Report.” 2014. OECD. http://www.oecd.org/std/microdata-access-final-report-OECD-2014.pdf.

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

Summers, Scott. 2017. “The General Data Protection Regulation (GDPR): Research and Archiving FAQs.” UK Data Service, UK Data Archive. https://www.ukdataservice.ac.uk/media/605048/summers-gdpr_faqs_final.pdf.