Deliverable Number: 6.8

Deliverable Title: Versioning requirements for curation and access to new forms of data

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

Deliverable type: Report

Dissemination status: Public

Submitted by: CESSDA (NSD)

Hervé L'Hours CESSDA (UKDA)

Darren Bell CESSDA (UKDA)

Sarah Butt ESS ERIC (City/HQ)

Gry Henriksen CESSDA (NSD)

Jindřich Krejčí CESSDA (CSDA)

Marianne Myhren CESSDA (NSD)

Janez Štebe CESSDA (ADP)

Øyvind Straume CESSDA (NSD)

Scott Summers CESSDA (UKDA)

Martin Vávra CESSDA (CSDA)

Date Submitted: August, 2018

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

www.seriss.eu @SERISS_EU

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

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

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

Please cite this deliverable as: L'Hours et al (2018) Versioning Requirements for Curation and Access to New Forms of Data. Deliverable 6.7 of the SERISS project funded under the *European Union's Horizon 2020 research and innovation programme* GA No: 654221. https://doi.org/10.5281/zenodo.1406217. Available at: www.seriss.eu/resources/deliverables

# Contents

# Deliverable Design, Structure and Content

This deliverable 6.8 follows on from D6.7 Generic high-level workflows for the curation of different forms of 'Big Data', by addressing the version requirements for appraisal of and access to new and novel forms of data as part of Task 6.3 (Connected curation and quality) of the SERISS WP 6: New forms of data: legal, ethical and quality issues. In keeping with the other T6.3 deliverables this text provides a broad analysis of the issues and challenges in designing and implementing a version management system and understanding the wider ecosystem of versioning approaches. These support the future application of versioning strategies to high level repository workflows.

## A note on Common Appendices

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

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

# Introduction

All data stewards, as actors in the research lifecycle with a role in creating, storing, managing and using data, have experience of different approaches to digital object versioning. Repositories, particularly those seeking to be 'trusted digital repositories' (TDR), like disciplinary data archives are systems which support the deposit, curation, access and reuse of data for a defined community of users. The digital objects they provide access to must be usable and understandable to that 'designated' community. Archives must understand changes in the technical requirements and the knowledge base of the community and update digital objects accordingly.

Temporal factors are critical to archives as they provide an infrastructure of people (with skills), workflows (business processes) and technologies all intended to provide a framework for implementing managed change over time.

Versioning is a set of mechanisms allowing us to uniquely identify the states of a digital object during its lifecycle, either at meaningful points in time (snapshots) or as an exhaustive history of incremental operations on the object, such that we can usefully isolate and identify changes to the object between adjacent or non-adjacent states. In cases where no part of the digital object has been discarded over time versioning also enables "rollback", such that we can revert the newest state of an object to a previous state. The concept of versioning data is inseparable from the management of digital objects and the wider infrastructure of

[1] https://eoscpilot.eu/

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

people, processes and technology which support curation.

Provenance is the process of capturing and representing a complete account of those changes, referred to as a provenance "chain". "Data provenance management is vital to science and scholarship, providing answers to common questions researchers and institutions pose when sharing and exchanging data".[3] A provenance chain usually includes state and temporal information.

From the earliest stages of a research, governmental or commercial activity multiple copies and different, changed versions of data and associated information are generated. "We may regard a new version to be created when there is a change in the structure, contents, or condition of the resource. New versions are created when errors that occur after data clean-up during data analysis are subsequently corrected; when the data are processed for the purpose of analysis; or when new data or data from other sources are added."(Krejčí, 2014) Granda & Blasczyk (2016) recommend that data managers "Implement a system of version control to maintain older versions of important data and documentation files. Users should be able to follow the changes made from one version to the next. Version control is necessary for users to replicate previous analysis or to test analysis done by others"[4].

Repositories may prepare both sensitive and anonymised versions of data from the same source under different access conditions at the same time. Researchers are expected and required to identify and cite datasets used as a research input to support reproducibility and trustworthiness. Data citation ensures that it is possible to uniquely identify[5] and retrieve the exact resource that the author used when answering the research question[6]. A version system which distinguishes between individual versions and keeps track of the differences also support the application of data authenticity and integrity measures to avoid or correct unauthorised modification of files or the loss of information[7]

These drivers for version control, which are common features of research data management training to researchers, are all important and valid. Applying a version system in a single environment with a narrow range of files for a limited audience is possible in even a primarily manual (non-machine actionable) curation situation. But as the types of data, range of supporting infrastructures and scope of the audience increase it becomes necessary to take a more rigorous approach which also makes a distinction between "versions" as the lineage of a single object and the looser usage of "versions" as alternative, modified instances of the same data for different purposes such as access.

For traditional and New and novel forms of data (NNfD) the design and implementation of version schemes ideally takes into account the needs of all human actors and machine-agents involved in the (research) data lifecycle.

NNfD are not so substantially different from 'traditional' data that past practices cease to be relevant, but they do present a set of challenges and opportunities which we cover in this paper. For the foreseeable future most repositories that engage with NNfD will manage a 'mixed ecology' of more traditional 'study-like' data alongside these emerging data types (L'Hours et al., 2018) (L'Hours et al, 2018a). The data must be evaluated alongside the new data sources, new technologies and the new research methods and opportunities they are

---

[3] Research Data Provenance Interest Group. 2017. Provenance WG Case Statement - DRAFT. Research Data Alliance. https://github.com/RDAProvIG/WGplanning/blob/master/CaseContent/Charter.md

[4] http://ccsg.isr.umich.edu/index.php/chapters/data-dissemination

[5] Data on the Web Best Practices, W3C Recommendation 31 January 2017 https://www.w3.org/TR/dwbp/#dataVersioning

[6] http://www.dcc.ac.uk/resources/how-guides/cite-datasets#sec:versions

[7] Krejčí, Jindřich, Johana Chylíková. 2017. Data authenticity in Expert Tour Guide on Data Management https://www.cessda.eu/Research-Infrastructure/Training/Expert-Tour-Guide-on-Data-Management/3.-Process/Data-authenticity

associated with.

The notion of using versioning to identify a change to a digital object implies a priori the notion of a digital object, and a common vision of what things are to be treated as a digital object for some purpose. The simplest digital object model describes the structure and content characteristics of:

- The data: the originally created or collected data points which are the subject of research
- The metadata: data about the data structured according to some formal standard
- The documentation: related contextual prose information of a less obviously structured nature

The concept of versioning is dependent on the environment and use of the digital objects. The aim is to design an approach to defining and managing digital objects (including their identification, versioning and citation) which supports services to data users and depositors alongside repository management of those objects.

It is important to note that versioning approaches which do not align across the lifecycle are not necessarily problematic or 'wrong'. While increased alignment and interoperability is always desirable, different versioning approaches may work with different parts of the lifecycle and some version models, such as those implemented by commercial software suppliers, social media platforms or government departments are unlikely to change in response to researcher or repository needs. Therefore, an understanding of the function and purpose of the version models in place and a documented approach to handling variation is important to manage changing digital objects across actors and workflows.

The section on evolving ecosystems below outlines some of the emerging activities which will influence the management and versioning of NNfD, particularly in Europe, over the coming years. We define the underlying concepts for defining digital objects, range of technical environments to be addressed, audience for versions and the actors involved and some legal and ethical considerations. The need to define clear levels of granularity for version models addressing different audiences is presented and the increasingly complex infrastructure of objects and business information which must be versioned is described. The implications of different data sources, types and characteristics, including truly novel challenges such as streaming data, are reviewed before we touch on data linkage and preservation. We conclude with some practical considerations for designing and implementing a version scheme.

## Evolving Ecosystems

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

These developments are strongly integrated with an identified need for audit and certification of data and services (including repository services) in line with the 'core' model (CoreTrustSeal, 2016) provided by the CoreTrustSeal as a baseline. Generic high-level workflows for the curation of different forms of 'Big Data' (L'Hours et al., 2018) expands on the notion of a 'core' Trustworthy Digital Repository (TDR) approach with higher levels of TDR process rigour and requirements granularity offered by Audit and Certification of Trustworthy Digital Repositories Standard (ISO16363[8]) and Criteria for Trustworthy Digital

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

Archives (DIN31644[9]). Both data and services must be, and be seen to be, trustworthy and FAIR (Findable, Accessible, Interoperable and Reusable).

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

## Digital Objects Defined

A digital object is an agreed boundary around a set of bits, ones and zeros, which allows us to structure, render and manage data and information. Some digital objects are self-contained file formats to render text, images and audio/video. These may be 'essence' formats that contain only audio/visual or graphics data. Other formats are 'wrapper based' which may also contain associated metadata and other information such as closed captions synchronised to the spoken word (Kovalick, 2012) These wrappers function like 'zip' file formats though research data management environments may treat a wrapper format (such as MOV-QuickTime) as a single object while treating the components of a zip file as independent entities. Knowing whether a file or bitstream has been be encoded with compression, encryption, etc., or bundled with other files or bitstreams into larger packages is addressed by the PREMIS Data Dictionary as a 'Composition Level', noting that:

> When multiple file objects are bundled together as filestreams within a package file object (e.g., a ZIP file), the individual filestream objects are not composition levels of the package file object. They should be considered separate objects, each with their own composition levels. For example, two encrypted files zipped together and stored in an archive as one file object would be described as three separate objects, each with its own associated metadata (PREMIS Editorial Committee, 2015)("PREMIS Data Dictionary for Preservation Metadata, Version 3.0," 2015).

Whether a workflow addresses one or more digital objects may depend on the rights model of the underlying formats. The image database and editing tool Adobe Lightroom uses XMP format [10] (Adobe, 2015) sidecar files to record change actions for proprietary 'raw' file formats while for DNG, JPEG, PSD, and TIFF formats the changes can be written directly to the file.

Different actors may have different technical and use perspectives on a digital object, and on what construes a digital object (see Perspectives on Digital Objects below). In addition, the schema defining the structure and permitted content of the format may be versioned over time to function in new rendering environments (such as Microsoft Word software support for .doc and subsequent .docx formats). These schemas are a dependency for object management and are themselves also subject to versioning.

For research data management and repository workflows, files are grouped and managed using a higher level logical collection of objects e.g. a 'data collection event', a 'study', or a 'wave'. Because these are logical rather than physical structures, designing or designating objects is (somewhat) arbitrary depending on local user expectations and technical needs.

Each wave of a study could be defined as:

- a different object with metadata to say "has previous version" or

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

[10] http://blogs.adobe.com/jkost/tag/xmp-sidecar-files

- versions of the same object

Or data, metadata and documentation could be defined as:

- separate objects,
- or as a single 'study'

The OAIS ("Reference Mode for an Open Archival Information System (OAIS)," 2012) designates SIP (submission), AIP (archival) and DIP (Dissemination) information packages as objects for archival workflows, but these may not reflect a repository users' understanding of the objects they deposit or access. It is important to distinguish between a single 'complex' object which has multiple parts and a collection of discrete objects which have specific relationships within the collection. Once an object is defined there remains a need to consider what 'parts' of an object will be addressed by the curation process (study as a whole vs granular structural metadata) before other key issues such as identification and citation (see sections below) can be addressed.

- Groups of objects e.g. objects for a particular project, topic, funder or date range
- Parts of objects e.g. questions or variables or file

As for the agreed boundary of an object these group/part designations are may vary across different curation environments must be designed to serve some purpose and to consider whether groups and parts of objects may be overlapping.

Until there is a common vision on these digital objects it is not possible to define their change management and the associated versioning policies.

## Perspectives on Digital Objects

All copies (see copy management below) and instances of digital objects must be understood and aligned within a versioning system and there are a number of academic and technical perspectives to support us in thinking about object definitions.

The Functional Requirements for Bibliographic Records (FRBR)[11] from the IFLA[12] supports resource discovery and access through the following conceptual entities: A Work realized through an Expression embodies in a Manifestation exemplified by an Item. These concepts are used to support an understanding of the differences between an intellectual product and the different physical and digital instances which may derive from it such as revisions, editions, translations etc. (Tillet, 2004).

The PREMIS metadata model[13] addresses four different types of object: a file (in the traditional sense) a bitstream (a sequence of bits which are subsets of files), a representation (all the files needed to render an intellectual entity) and the parent intellectual entity itself (Caplan, 2017). It must be clear within the version model how these parts influence each other and whether they must be versioned together or separately.

An archive may receive a single deposit at the end of a project but generate different dissemination packages for different purposes, e.g. an anonymised and redacted versions for Open Access and a version retaining sensitive personal data via a local or remote safe access environment. Decisions must be taken as to whether ongoing changes (additions,

[11] https://www.oclc.org/research/activities/frbr.html

[12] International Federation of Library Associations and Institutions: https://www.ifla.org/

[13] https://www.loc.gov/standards/premis/

corrections, deletions) need to be synchronised across all available instances.

The intellectual entity of a wave of a longitudinal survey is defined in terms of the individual file (data, metadata, documentation) and may define further structural or other metadata to the bitstream level.

As addressed in the Technical Environment below, file and database management environments contrast with datastores/linked open data (LOD) approaches which natively support linkage of data within and across multiple source objects. A range of data products may be derived through data linkage, whether these need to be versioned depends on whether the repository may decide to discard them, retain them for a period, or preserve them for the long term (see Preservation, below).

### Copy Management

A version model must also consider the location and number of copies of a digital object and the level of control they exert over those copies.

Statistical files for dissemination to researchers may be released as a single user-facing version with instances in multiple formats (SPSS[14], SAS, and R[15] etc.) across multiple environments (file server, Nesstar[16] SQL back end, Hadoop[17] triple store). Additional copies may also exist within software development and production environments.

Best practices for data storage indicate a multi-copy redundancy solution where data may be copied to on-site, near-site and off-site locations on a variety of media (SSD, spinning hard drives, tape). In addition to maintaining integrity across copies it is important that all copies are addressed by the version model and clear back up periodicity must be defined to identify the level of any possible data loss in the event of a disaster. For example, if a file changes daily but is backed up only weekly, there is a potential risk that only partial restoration of the data will be possible. In this interim period the repository is managing two non-synchronised copies of the digital object.

Data which is not curated through a series of trusted systems (across the data lifecycle), or which is shared ‘ad hoc’ with no records being kept can lead to uncontrolled copies of data ‘in the wild’. One challenge of object, version and copy management is to make it easy for researchers to share data openly while maintaining the necessary identification and location management.

## Technical Environments for Version Management

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 of the new features of NNfD, including the re-tooling of any technical infrastructure. The primary ‘traditional’ target object addressed by SERISS is ‘the Survey’, as either quantitative microdata, aggregated data, or qualitative data. The ‘survey’ usually implies data, metadata and documentation as the simplest object model. This must be managed over time and, for longitudinal surveys, through multiple connected waves of data collection. Establishing relevant features of the objects helps us define what metadata we need to manage through the lifecycle. Such metadata encompasses file formats (and dependent technical environments necessary to render

[14] https://www.ibm.com/analytics/spss-statistics-software

[15] https://www.r-project.org/about.html

[16] http://www.nesstar.com/

[17] http://hadoop.apache.org/

them), clear time-stamping of actions, versioning, change logs and structural metadata. These allow us to relate originally deposited digital objects to the later versions of those objects designed for dissemination to researchers or designed for long term digital preservation. NNfD differ from the 'standard' survey/aggregated statistics or qualitative data which the CESSDA organisations typically hold in a number of ways (adapted from L'Hours et al, 2018a (L'Hours et al., 2018)).

The versioning model must take account of the technical environment in which the digital objects will be managed. Many repositories will work with a mixed ecology of systems. If an object is managed in more than one environment at a time e.g. where new forms of data are managed both as traditional files and as data processed (flattened) into a "datastore" (see below) during all or part of the curation or access processes, this must be considered when designing versioning models.

With the exception of the datastore paradigm the underlying technologies to natively support version identification and management such as Version Control Systems (VCS), Content Management Systems (CMS), Wikis and word processors have existed for many years (Allsop, Somerville and Shipsey, 2007)(Allsop & Shipsey, 2007).

## File-driven Environments

Most version advice related to research data management (RDM) for the pre-repository phase of the lifecycle refers to file-driven workflows[18]. The majority of version guidance for archiving also focuses on file-driven technical environments with clearly bounded digital objects (LSE, 2008).Recommendations cover standardised file naming, directory (or folder) naming with directories structured according to good 'information architecture' practice. Access and edit permissions are controlled via standard file-based authorisation mechanisms (e.g. NTFS Access Control Lists in Windows, or POSIX permissions in Linux) and authentication systems such as Microsoft Active Directory (AD).

These file-based systems tend to support access modes where the full digital object is discoverable through a metadata catalogue and where access requests usually result in a direct download to the users' local machine for further analysis.

## Databases

To support presentation, visualisation and analysis of data at a more granular level, often through web-accessible interfaces, data may be integrated into SQL or other relational database protocols. Common examples are environments that support addressing research data at the question or variable level. Despite the verbose logging offered by most database products, these are seldom human readable (or at least readily digestible) so some consideration of the audience requirements for versions is necessary.

## Datastores: RDF Data Lakes and Graph Databases

Most data curators and users are familiar with file-based systems and have an awareness of at least the front-end interaction with database-driven systems, even if they don't work with the underlying data structures. The emergence of linked data formats such as RDF, or NoSQL formats like key-value pairs, and new technical environments to support their use at scale (such as Neo4J[19] or Hadoop) present a number of opportunities for research data analysis and management but also imply a degree of re-skilling among researchers and repository staff as we go beyond the traditional paradigm of clearly defined sets of files

[18] https://www.cessda.eu/Research-Infrastructure/Training/Expert-Tour-Guide-on-Data-Management/3.-Process/Data-authenticity

[19] https://neo4j.com/

moved together from one storage location to another.

> 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 previously modelled/structured, while a data lake stores it all—structured, semi-structured, and unstructured. (L’Hours et al, 2018)(L’Hours et al., 2018)

See Data Linkage Modes

## Audiences, Actors and Agents

A broad understanding of the wider stakeholder ecosystem (legislators, standards bodies, funders, vendors) and the local organisational context (accepted file formats, software products in toolchain, depositor and user expectations) is necessary to define a usable version model. These actors' should also be assessed in the context of whether they precede or follow the repository in the research data lifecycle. NNFD expand the range of actors that repositories and researchers interact with and make it even more important that roles in research-data handling are defined in advance (OECD, 2016).

It is important to define the audience(s) for a version model, and to identify and meet their needs. These may be both human actors and machine agents.

Data depositors working with longitudinal/multi-waves of data will ideally agree a submission schedule (PAIMAS, 2004) with the repository that supports both planned deposits and ad hoc deposits (additions, corrections etc.).

Data repositories will undertake numerous curation actions on deposited data to generate dissemination and preservation versions. Not all of these versions/version metadata will be relevant or useful to depositors or data users (see granularity below).

Data users will seek a consistent and clear approach to understanding objects and versions and to communicating the changes and their significance between versions (“what has changed?”, “why?”, “will this impact my analysis”).

Repositories define a ‘designated community’ of users whose knowledge base and technical requirements are catered to over time the designated community may change their preferred technologies over time necessitating forward migration of digital objects to new formats or emulation environments. The scope of the designated community may change (e.g. from serving social scientists to supporting those working at the intersection of social science and medicine). This will imply changes to the descriptive documentation of the data to provide context to a wider range of researchers. Data may be accessed by a community of users which is wider than the designated community, but the designated community should be the primary targeted audience for version models.

Legislation, ethical standards and technical standards may all change (be versioned) with little or no influence from the curators (see Distributed and Dependant Objects below). These changes may require new versions of curated objects.

The audience of versioned data may include machine-agents. Machine-harvesters (e.g. OAI-PMH[20]) or API users will need query/response mechanisms to limit data collection to only those digital objects which have changed, usually embedding this information in a query for ‘pull’ or in accompanying metadata for ‘push’ (Rauber & Asmi, 2016). Alternately a

[20] https://www.openarchives.org/pmh/

consuming application may run a comparison of the current and prior versions[21] to identify changes.

Alongside these 'consumers' of version information we also need to define which actors have custody over the digital objects at which points in the lifecycle and their level of rights over the objects (permission for new version generation).

## Legal Considerations

For datasets which include personal data the concept of 'data protection by design and default'[22] must be built in. Modern datastores provide the ability to link structured and less structured data from multiple sources at scale and in these environments even source datasets have been through an anonymisation process may be at risk of re-identification.

While the emergence of the General Data Protection Regulation (GDPR) has directed attention to the storage and processing of personal data, these personally identifiable data points may exist across a range of systems, environments, database tables and files. This distribution of elements of personal data across multiple parts of multiple systems may complicate data protection. When such systems are redesigned the concept of defining an individuals' records as an addressable, versioned object simplifies personal data management. This includes cases where a data subject withdrawals consent, to implement the 'right to be forgotten, or to enact record deletion after an agreed period of retention[23]

The models and actions must also be appropriately documented and communicated to provide evidence for what the GDPR refers to as 'appropriate technical and organisations measures'. Requirements under the GDPR to provide clarity on data processing means that more granular and well-documented version control may be required.

Intellectual Property Rights (IPR) such as the terms of use of social media data may also impact object and version models (see Distributed and Dependant Objects, and Social Media sections below).

## Versions and Granularity

Ideal archival practice involves always working on a copy of the data rather than the original but for those working with research data at massive scales this is not always possible. The generation of 1PB[24] data generated per second at CERN[25] requires that a large proportion of the data is not retained. Even much smaller research entities will encounter situations were retaining all data, or working on multiply redundant copies is not practical. In these cases the version process is not reversible (a later version cannot always be 'rolled back' to an earlier state) and the decisions and actions taken between versions must be even more carefully logged.

- For files at a smaller scale it is often possible to retain an untouched original, but in manual processing work on files there is not automated granular versioning and new instances may only be saved at the users' discretion when saving a new version or

[21] https://en.wikipedia.org/wiki/Diff

[22] https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-by-design-and-default/

[23] https://www.iso.org/standard/62542.html

[24] https://www.techrepublic.com/article/cern-we-generate-1pb-each-second-heres-what-thats-taught-us-about-big-data/

[25] https://home.cern/about

checking files back into a managed VCS.

- Data held in relational database may be more easily managed through multiple versions, but full logging of every change event may not be human readable.

- A graph database may support granular versioning at data-point level but will still depend on the notion of a parent 'object' so that changes can be 'rolled up' and managed at object level.

The RDA Working Group on Dynamic Data Citation (WGDC)[26] notes the challenges of storing all revision versions in large data scenarios and notes that from a data citation perspective not all objects, questions and variables across every historical version may be citable (Rauber & Asmi, 2016).

### User Experience of versions

Version model design should be developed with different audiences in mind and version. Associated change metadata should be delivered at a level which meets the audiences' needs.

Full database logging may support machine actionability without human readability. Archives' internal change logs may be granular and verbose. Changes presented to end users should be simple and clear while providing access to more granular information as required.

The presentation of data through web interfaces widens the possibilities for communicating version changes to users. We can move from major/minor version change logs to providing contextual information on change e.g. versions may be tagged to the history of a question/variable or visualisations could be presented to let users compare and contrast two different versions of a dataset.

## Distributed, co-dependent infrastructures and objects

### Distributed Infrastructures

Repositories often work in complex partnership environments with a variety of actors. With the emergence of efforts such as the EOSC[27] the number of actors and entities will grow and the possible copies and versions of a dataset will expand. Identifying and tracking these copies and aligning versioning practices will be key challenges in managing workflows across the research lifecycle. The proposed and evolving 'rules of participation' envision certified FAIR datasets in certified Trustworthy repositories (TDR); this implies that the object characteristics and environment characteristics will be closely aligned in workflow management. A FAIR object with full provenance managed by a series of TDRs will ideally have an unbroken chain of version history.

With the need to rely on an increasing number of infrastructure partners and an increasingly large range of data providers and sources, including social media platforms, there is a need to identify, manage and formalise relationships between a wider ranges of stakeholders. To assure demonstrably trustworthy relationships and service provision the range of contractual and service level agreements in place will need to be more transparent. For social media platforms and other commercial providers/sources of data there are few established relationships with archives. Commercial organisations may change the terms and conditions of membership and of data access and use with limited consultation and warning (see references to Twitter terms below). It will be challenging and resource intensive for all

[26] https://www.rd-alliance.org/groups/data-citation-wg.html

[27] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

repositories holding social media data to stay up to date with changing conditions which could impact their right to store, curate or offer access to data. Coordinated registries of these terms and conditions accessible to all research lifecycle actors would provide a valuable service by clarifying legal terms and reducing duplication of effort.

## Distributed and Dependent Objects

Historically repositories have held and managed the primary data, metadata and documentation in their entirety but numerous complex partnerships of entities underpin our understanding of the outputs that repositories make available, and not all of these can be stored alongside the data. Data managers already rely on XML schemas, standards and controlled vocabularies/ontologies managed by third parties and the emergence of linked open data implies that distribution is built into the data model.

Terms and conditions for social media data may be challenging for storage and sharing of data for research purposes  Twitter's terms of use, for instance restrict a researcher to depositing and sharing tweet identifiers. These 'dehydrated' identifiers must be rehydrated by reference to the Twitter API further research these must be. As deleted tweets or the content of deleted accounts will not form part of the rehydrated dataset secondary use may be based on incomplete data/metadata. In this case a repository is only maintaining a part of a distributed digital object (see Social Media data, below).

The OAIS model defines 'representation information' as the information needed to render the 'content information' (the information to be preserved). But in real-world practice a dataset of JPG images will not be accompanied by instructions on how software converts bits to pixels, and will often assume that accessing JPG rendering software is part of the users' knowledge base. But this will not always be the case for more obscure file formats and the National Archives PRONOM[28] provides a registry of software products, and the file formats which each product can read and writes to support these.

The provision of such registry services offering human and machine-readable access to research-supporting information can greatly reduce costs and duplication of effort. This centralised service model can be extended to information which is critical to support object metadata such as controlled vocabularies and ontologies. Registries of data types,[29] standards,[30] researchers,[31] repositories[32] and identifiers (see Identifiers and Citation below) are all emerging to provide common reference services to data infrastructures.

As well as providing benefits to data repositories the registry model presents a number of risks. It can be challenging for registries to move from project funding to an operational business model and the FAIR data high level expert group acknowledges that Registry services need to be expanded in scope and scale (Hodson et al., 2018)(Hodson, 2018). The repository must trust the registry provider and most of the TDR standards applicable to data repositories apply equally to these 'metadata repositories'.[33] The repository must understand the version model of the registry in terms of both the underlying schemas and the provision of updates. The repository must decide how and when to adopt updates into local systems and define when these changes require new versions of dependent digital objects. A user registration form needs to contain only current countries from ISO-31661, while a controlled vocabulary covering historical datasets on Burma must take a decision on which resources

[28] https://www.nationalarchives.gov.uk/PRONOM/

[29] http://typeregistry.org/registrar/

[30] https://fairsharing.org/

[31] https://orcid.org/

[32] https://www.re3data.org/

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

are updated to reference Myanmar. In both cases changes must be identified, adopted and the implications for object versioning managed.

As the number of dependent information sources expand, the repository must also consider whether their needs can be met by:

- Linking to registry providers and retrieving records in real time – this can have consequences for the integrity and versioning of a digital object when an externally referenced object is updated,
- Retrieving records and storing them locally with the digital object- this can cause these records to become de-synchronised from the originals
- Copying registries or part of registries (e.g. CVs) to local environments periodically and managing the adoption of new versions.

See also early/late binding under *QVDB, QDDT and DDI* below.

### Supporting Documentation: Business Information

One class of information which may be treated as a dependent part of the wider object model is the business information which governs how data is curated and the outcomes of those curation actions. Policies and procedures for archiving processes have important evidential value and a given procedure should be time-stamped for its period of validity and linkable to the datasets it is applied to. Versions of datasets and the rules to be followed for their management must be aligned and preserved over time. This information which may be locally developed, disciplinary, or derived from a general purpose registry (see above), can be divided into three broad types:

**Prescriptive** information governing the legal, ethical and process requirements, whether at a higher policy level or at a granular operational level as a workflow or procedure.

**Reference** information providing 'look ups' of preservation file formats, open source software, Trustworthy Repositories or researcher identifiers.

**Transactional** information describing the outcome of a curation or other repository actions such as "identity validated", "virus check passed" or "QA check failed". This information is always directly associated with an identified curation action which has been prescribed (e.g. as part of a procedure) and may make use of reference information (e.g. a virus database).

## Metadata, Persistent Identifiers and Citation

The use of standardised metadata across technical environments is critical to ensuring the free movement of data without compromising data governance. But these standards are also versioned and must be aligned with locally held data and systems.

Dependent objects (see above) supported by metadata standards include research entities,[34] funders,[35] and researchers.[36]. Repositories depend on a variety of metadata systems for data description and management. Social science data, including surveys, is increasingly aligning on the use of the DDI standard but also dependant on a wide range of other structured metadata standards including PREMIS, METS and DC. These last three make schemas available via the Library of Congress and have generally aligned on a

[34] https://www.eurocris.org/cerif/main-features-cerif

[35] https://www.crossref.org/services/funder-registry/

[36] https://orcid.org/

version model which focuses on the presence or absence of backwards compatibility. Minor revisions are backwards compatible, major revisions are not.

> "This revision is backwards compatible to version 3.6 and therefore only includes changes that do not result in invalidating existing MODS records." [37]

Locally applied disciplinary or domain-specific metadata may also be transformed (and often simplified) for other environments such as aggregated catalogues[38] which will version their own underlying schema.

Citation of data, metadata or publications is dependent on a clear object model which allows us to 'identify' an object and a clear version model to ensure the citation refers to a version of the object in a defined state.

It is good practice for identifiers to be:

- unique (within a defined domain)
- resolvable (supporting the user in getting to the cited object)
- persistent (remaining unique and resolvable over time)

A citation is, on the other hand, is a string of metadata values in a standard sequence which is used to minimally describe and enable discovery and retrieval of a resource. Best practice citations include version metadata and a persistent identifier resolvable to the resource, or at least a proxy for the resource such as the version history metadata[39].

A variety of persistent identifiers exist including Archival Resource Keys (ARKs), Persistent Uniform Resource Locators (PURLs) and Uniform Resource Names (URNs). Among CESSDA archives Digital Object Identifiers (DOIs) based on the Handel system[40] and supported by DataCite[41]. Under the expanded DataCite mandatory properties it is recommended that those minting DOIs distinguish between major versions (which trigger a new DOI) and minor version which are recorded within DOI metadata[42] .The same approach is suggested by the UK Data Service: data repositories should ensure that different major versions are independently citable with their own identifiers[43]. But in each case the decisions over which changes constitute a 'major' increment are left to local practice which should be clear to end users and ideally reflect common disciplinary or domain practice. At the GESIS[44] repository changes to data and documentation are logged at of three levels:

- Major changes - (as is addition/deletion of one or more variables or cases in a data set)
- Minor changes - (changes relevant to the meaning of a variable, e.g. recoding)

---

[37] https://www.loc.gov/standards/mods/mods-3-7-announcement.html

[38] https://www.europeandataportal.eu/

[39] https://discover.ukdataservice.ac.uk/catalogue/?sn=5050

[40] http://www.rfc-editor.org/rfc/rfc3650.txt

[41] https://www.datacite.org/

[42] Data Cite Metadata Working Group. (2017). Data Cite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.1. DataCite.V. 10.5438/0014.

[43] http://www.dcc.ac.uk/resources/how-guides/cite-datasets#sec:versions

[44] www.gesis.org

- Revisions - (e.g. in spelling of label)

With a resultant three tier approach to versioning: 2.6.5

## Dynamic Data, Streams & Queries

Not all data is at rest, and data infrastructures are now having to address “streaming” data. Streaming is a classification applied to data that surpasses an arbitrary threshold of ingest frequency (determined locally by the repository), such that it becomes impracticable to assign discrete version numbers for each fragment of data that is ingested, or doing so no longer serves a useful purpose. It is significant to repositories whether the data they are curating (and which users are querying against) is a snapshot of a streaming source, or something more dynamic.

Data may be added to, or removed from while in motion and routed through different dataflows. The underlying schema which describes/dictates the content of the stream may also be versioned. Repositories have yet to develop best practice on versioning for these use cases.

The Research Data Alliance (RDA) working group on dynamic data citation[45] suggests identifying the exact version of a subset as it was used during a specific execution of a work. If the data source is continuously evolving this can be achieved by ensuring that data sets are assigned persistent identifiers (PIDs) and that timestamped queries that can be re-executed against the timestamped data store (Rauber & Asmi, 2016).

DataCite recommendations for citing rapidly changing data proposes four possible approaches (DataCite Metadata Working Group, 2017):

a. Cite a specific slice or subset (the set of updates to the dataset made during a particular period of time or to a particular area of the dataset.

b. Cite a specific snap-shot (a copy of the entire dataset made at a specific time).

c. Cite the continuously updated dataset, but add an Access Date and Time to the citation.

d. Cite a query, time-stamped for re-execution against a versioned database.

The “slice,” “snap-shot” and "query" options require unique identifiers. Option (c) necessarily means that following the citation does not result in access to the resource as cited. This limits reproducibility of the work that uses this form of citation.

## Data Sources, Types and Characteristics

The level of past provenance at the point of data deposit in an infrastructure is a strong indicator of historical versioning practice. Version approaches can be consistently applied across a range of data types and characteristics.

### Administrative Data

Administrative data, with population data being a common example, is not consistently developed with linkage and re-use by third parties in mind. It may be prepared within strict internal governmental guidelines but not all steps of intermediate data design and processing are formally recorded or version-managed (or this may not be made available

[45] Working Group on Data Citation: Making Dynamic Data Citeable“. https://www.rd-alliance.org/group/data-citation-wg.html

upon deposit). The availability of an administrative data spine in countries like Norway (see below) provides opportunities for early alignment of version approaches between government departments and survey infrastructures. The recent ESRC Longitudinal data report (Davis-Kean et al., 2017) noted

> "An administrative data population spine on which to base current and new longitudinal studies – would provide a key underpinning and transformative element of the UK's research data infrastructure; its creation would involve a number of challenges but the returns would be significant and reach beyond social science."

### Norway, RAIRD technology and Microdata.no[46]

This section presents some of the work within the NSD - Norwegian Centre for Research Data (NSD[47]) to address the management and versioning of administrative data.

The size of Norway means their population datasets fall into a 'medium' data category of 10-100 million records, but these still call for dynamic applications to handle dissemination. There are two key areas of activity with some shared characteristics; the RAIRD software/microdata.no site described here and the ESS Cumulative Data Wizard (below).

A key challenge is to support citation and reference that goes beyond the static links to static datasets which are sufficient for more file/study driven workflows. There is a need to support citations that support filters applied to data sets with each selection containing at least one temporal dimension to ensure systems support a valid and meaningful citation. This goes beyond subsetting from a static object undertaken with traditional survey data analysis (e.g. subset of variable and subset of population) to deliver granularity well below the dataset/study/survey level.

For bigger and more dynamic data sources the solutions for citation haven't been finalised but citability has been designed into the platforms.

The traditional Nesstar approach strains the OAIS reference model as each 'deposit' overwrites the previous version.

[46] Meeting arranged by Gry Henriksen (WP6) with Hervé L'Hours (T6.3) and Ørnulf Risnes (Head of Software Development NSD) to discuss the NSD work as it related to the versioning, citation and persistent identification (PID) of new and novel forms of data and data research platforms.

[47] http://www.nsd.uib.no/nsd/english/index.html

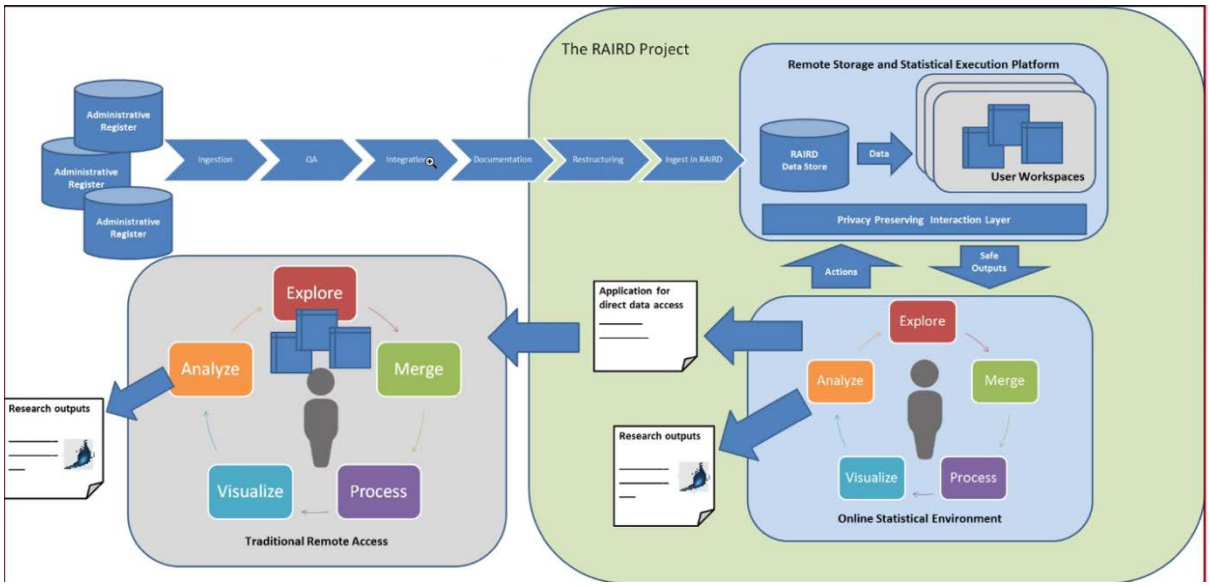


**Figure 1: Interactions with the RAIRD project**

**Norwegian Administrative Data Register**

This work is less survey focussed and more big data. Microdata.no is underpinned by the RAIRD technology project work which is SQL based, though the logical layer doesn't *need* to be addressing a SQL data source.

Each client session is working on a particular 'revision' of the datastore (of a particular underlying object). Any change of or removal of data generates a new revision as the implied contract is that each datastore revision is immutable. New client sessions will be connected to the latest revision. Older client sessions will be informed a new revision is available but will continue to be connected to the older revision unless they specifically request otherwise.

Client sessions provide for statistical analysis packages within the web browser as a remote online environment. Users can undertake standard activities such as running a script on two income variables, merging, comparing, subsetting and summary analytics. Each script is run against a defined revision and is citable forever with the same results guaranteed over time. The citation is effectively a self-contained recipe to re-run the query with the same parameters against the same data source.

The RAIRD approach does not fully duplicate (unchanged) for each revision but the user experience is analogous to working on a single AIP. One challenge in justifying the approach (e.g. to Stats Norway) was concern over a proliferation of versioning. The approach selected can be compared to git[48] version control as each commit generates a new version containing all historical changes. Each version can be referenced by 'name' but also presents business/public (externally visible) versions which include a given set of (internal) microversions. On 'publish' all changes are transparent to the end user.

This approach is supported by a parallel metadata database which is used to track new revisions. New revisions will retain unchanged SQL tables and generate and tag new SQL table for changes. The metadata database supports appropriate 'pointing' to unchanged and new tables in a way which provides clear versioned revision objects to end users. For most data archived at NSD Stats Norway undertakes data cleaning and preparation through scripts run in SAS-Oracle whose results are then deposited with NSD (depositors are encouraged to document their changes at the point of deposit) but in the case of

[48] https://git-scm.com/book/en/v2/Getting-Started-About-Version-Control

Microdata.no the data are retained in servers at Statistics Norway.

The microdata.no system currently holds personally identifiable data so there is no provision for downloading outputs, though the system could support more open data which permitted this. Currently all data management and analysis is undertaken internally and automatic SDC applied to any results before presentation to the user. Full data access is not permitted, but this limitation is a trade off against the very long processes that are required to otherwise negotiate access to the data. Queries against non-anonymised data sources are instantiated as scripts/recipes which can be used as citations and to reproduce the query (if the actor using the query has appropriate permissions to the dataset). This presents the challenge of disconnecting citation from access management (not an issue with open data).

The source data required a degree of data wrangling before its ready for research and there is the potential for offering this as a service within the system, which would then provide a more complete provenance trail of activities currently undertake pre-deposit by monitoring each microversion change, but still providing clear, meaningful public versions to users.

There is no need to justify a request for each variable and a user can download the analytical output, but not the raw data. There is no integration of less-sensitive data at this stage, but the product could be used to serve this.

Data is considered an exception to the need for consent-driven management as StatsNorway use them for the production of official statistics and there is a built-in assumption that this be shared for research. The issue of excisions in the case of a withdrawal of consent are therefore not an immediate issue but any such changes would be at odds with the version control regime.

Citations link to a landing page due to access restrictions.

Long term digital preservation for ESS remains under discussion but the move from file to SQL based provides two benefits:

1. All data is integrated (lossless) into the system

2. Supports preparation for data in a variety of formats and with bounded as a variety of subsets.

The underlying version logic is that all "tagged" versions of the combined metadata/data sets are equally available over time. The main server endpoint for data sets has a /revisions method, that lets consumer list out/navigate versions. The version-abstraction is more explicit than for traditional file-based data. Smaller/intermediate changes between "tagged" versions are available for internal management/curation systems to consume, but not tangible to outside users who only see the "tagged" versions. For microdata.no all access to data goes through a metadata-layer powered by Datomic[49] to keep track of both metadata and 'tags' providing git[50]-like capabilities in dealing with metadata changes and data changes as long as data are either embedded in metadata or metadata have "pointers" to various data versions e.g. file names, database table names.

## Survey Data

Survey data from survey ERICs or projects may have long-standing deposit relationships with archives and are subject to ethical and methodological review throughout their funding applications, development and implementation. As with Administrative data the full version

[49] https://www.datomic.com/

[50] https://git-scm.com/

history may not be available. For other data infrastructures surveys will be expected to provide assurances that the GDPR-compliant technical and organisational measures are in place to protect personal data. This may lead to closer relationships with repository infrastructures or surveys seeking TDR certification, in either case more granular and documented version management is likely to be required.

### Versioning of surveys: the ESS Example

ESS[51] provides both data and documentation with data following a general trend towards the DDI (Data Documentation Initiative) versioning. In common with many other business environments, documents have tended to receive less versioning attention than data. A brief overview of some of the main features of versioning within the ESS is given below.

**Documentation**

For a survey infrastructure there are three user groups for whom document versioning is important:

**End users** who access public documents on the website.

**Internal stakeholders** who access "final" versions of survey documents via the ESS intranet in order to implement the survey e.g. copies of the questionnaire for translation.

Documents made available to both of these groups are dated and, if changes are necessary following publication, a change log detailing any major changes will be included at the front of version 2 (and any subsequent versions) of the document. Internal stakeholders are sent an email alert in the case of major changes.

**Version control of documents in production** which may be being worked on by multiple stakeholders. This presents issues of conflicting edits and version control familiar to many businesses, but perhaps this is a particular issue for cross-national survey infrastructures collaborating across institutions in multiple countries. Various technological solutions such as Redmine or Googledocs - as well as the Survey Management Portal (SMaP) being developed under SERSS workpackage 4 – and which allow for document sharing and collaborative working are being explored to address this challenge.

**Data**

With respect to versioning data, for each survey round ESS release edition 1.0 of the data which contains data for the majority of ESS countries who completed fieldwork and deposited data in time for a release in the November following fieldwork. Following the initial release, ESS data is regularly updated (up to twice a year) to include more countries and address any errors in the data found since last release. If an update adds data from new countries, the major edition number is incremented (e.g. from 1.0 to 2.0), whereas if an update only contains data corrections, the minor number is incremented (e.g. from 2.0 to 2.1). Edition 2.0 is released around 6 month after the initial release and contains data for all, or most, ESS countries in a particular round. All changes in the data made in an update are documented in the version notes.[52] Major alerts/updates are announced to users by email and are available in the alerts archive.[53] In the near future, ESS will attach persistent

[51] http://www.europeansocialsurvey.org/

[52] http://www.europeansocialsurvey.org/data/ESS7_version_notes.html

[53] http://www.europeansocialsurvey.org/data/alerts_archive.html

identifiers in the form of DOIs to each round of the survey.

The ESS Cumulative Data Wizard[54] gives access to cumulative data from countries that have been included in the integrated ESS files in two or more rounds. It addresses a *.nesstar file through nesstar file as a service (NFaaS), sub-setting common variables across nine waves, and across rounds by nation, generating a bespoke 'dissemination information package' (DIP) from the underlying API (Archival Information Package). The data file underlying the wizard is editioned the same way as other ESS data – major number incremented when adding countries, minor number incremented when only updating data. When data from a new round is added to the wizard, it is regarded as edition 1.0 of a new file, rather than an update (i.e. "ESS1-7 ed. 1.0" will become "ESS1-8 ed. 1.0" when ESS8 is added). The cumulative data is in general updated once every 2 years. Since the cumulative data is created by merging data from all ESS rounds, the editions of each rounds' data is documented on the wizard's web page and in each generated file.

**QVDB, QDDT and DDI**

Survey infrastructures are increasingly making use of the DDI[55] to version survey metadata (including question items and variables) though the standard has no prescriptive model for versioning, only controlled attributes to allow data managers to describe their own versioning system and policy. Several tools being developed in the SERISS[56] project and elsewhere use the DDI standard and benefit from its versioning schemes. Though this is not yet standard practice adoption is being community-driven, including by ICPSR[57]. Two tools being adopted by the ESS which make use of DDI 3.2 (and follow general DDI versioning rules as described in in chapter 3.0, p. 10 and onward in the DDI Part I Technical Document[58]) are the QDDT and QVDB.

The Questionnaire Design and Documentation Tool (QDDT)[59] provides considerable flexibility in versioning, recognising the complexity of the questionnaire design process. It supports major and minor version changes as well as saving of 'work in progress' revisions. The rationale for change is documented in a standardised way for each minor (typographical/orthographical) change and major change (of meaning). A standard major/minor digit (N.n) nomenclature is used and major changes may be classified as: conceptual improvements, real life changes, additional content elements or 'other'. Comparison functionality lets users compare values pre- and post-change. In-progress changes are flagged in the user interface with yellow exclamation marks while all its parent elements display a green exclamation mark.

Lower order, highly reusable, elements like Categories, ResponseDomains, QuestionItems and QuestionConstructs are early bound in the tool. This means that when an element is referenced by another element, a specific version of the referenced element is always retrieved. This allows the user to select and bind a specific version of a ResponseDomain to a specific version of a QuestionItem, for example.

The Question Variable Database (QVDB)[60] provides full support for item versioning.

---

[54] https://www.europeansocialsurvey.org/downloadwizard/

[55] *http://www.ddialliance.org/*

[56] https://seriss.eu/

[57] https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter3docs.html

[58] http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/HighLevelDocumentation/DDI_Part_I_TechnicalDocument.pdf

[59] https://github.com/DASISH/qddt-client/wiki

[60] https://github.com/DASISH/QVDB

Versionable items in the QDVD are, for example, question items, variables, code sets, categories, data files and studies.

The QVDB repository uses a write once scheme to ensure provenance tracking and data protection. Each change to an administered item therefore increases its version number within the system (only one version level). Every version of an item committed to the repository is saved, allowing clients to retrieve a full version history of any item in the data base. Each version includes additional information: The authenticated user who committed the version, the date and time the version was committed and an optional message describing the reason for the version Tagging. A specific version of any item or set of items may be tagged. A tag is simply a name given to the version so it can be easily referenced in the future. Tagging can be used to mark milestones such as publication.

### Social Media Data

Social media data is created to deliver services which may be added, changed or removed rapidly with little notice, in response to changes in commercial focus, user demand and public perception (See L'Hours et al, 2018b for the implications for Appraisal and Selection). The schemas and data are not created with research in mind, but to be performant at scale. Underlying data schemas are not transparent and legal data access modes (where they exist) may not support consistent repeatable querying or selection of data as representative of a particular population. Terms and condition may also change, presenting a challenge to maintaining compliance when data has been deposited in an archive (see references to Twitter above in Distributed and Dependent Objects).

Across the APIs available for social media platforms there may be limited control over the ability to consistently draw from the same sample. For example, experimental work by the GESIS archive on Twitter data experienced undocumented changes to the behaviours of geotagging metadata during the period of data collections (Kinder-Kurlanda, Weller, Zenk-Möltgen, Pfeffer, & Morstatter, 2017) (Kinder-Kurlanda et al, 2017). These factors and the need to manage ongoing changes to terms and conditions of access and re-use present serious challenges to clear version management and communication.

### Personal Data

As referred to under Legal & Ethical considerations (above) the presence of personal data in deposits or in repository-mediated linked data environments is a critical characteristic that influences digital object design and versioning.

## Data Linkage Modes

Data Linkage undertaken prior to deposit has less of an impact on repository decisions as the derived data products can be addressed in the same way as most data offers at the point of appraisal and selection and consent and IPR issues in the linked data may have been clarified. But the absence of fully versioned provenance (from the administrative, survey and social media data sources) reduces the quality and confidence in the overall dataset.

The main impact of linkage on appraisal and selection is if deposited data will be subsequently linked to other data in an environment controlled by the repository. In this case the version model will need to take account of:

- temporary data products generated during analysis
- those selected for submission for statistical disclosure control
- final approved outputs that are allowed to leave the repository-controlled system

Some linked data products may be curated by the repository and made available to researchers, and products generated by researchers themselves may prove more generally useful and be retained. All of these must be associated with their original source digital objects.

### Linked Data & Rights Management

Each digital object which forms part of a linked data product may have intellectual property and other rights, restrictions and obligations attached. Explicit or implicit consent from data subjects as to whether such linkages are permissible may need to be considered unless some other appropriate derogation is in place. The versioning model must support the tracking and integration of these from source objects. Compound IP in linked data may be challenging to define and communicate, and access criteria (including restrictions to protect personal data) may need to default to the lowest common denominator of permissions.

## Preservation

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

Similar considerations can be made regarding 'documentation' that is referenced in metadata. E.g. links to web resources and consequences of broken links to consistency of documentation...

## Designing and implementing a version scheme for NNfD

This section considers the steps necessary to design, and apply a version scheme in the sequential repository phases of the (research) data lifecycle within the context of the issues outlined above.

Version management is dependent on having a clear object model. Without a common understandings of the objects to be managed it becomes very problematic to consistently describe changes to digital objects, whether as new versions or as new copies.

Accurate version management is critical for persistent identification and citation of digital objects at any level of abstraction, from projects and studies, to files, to granular questions and variables.

It should be clear to all parties whether there are any phases of curation which involve the loss of data or an inability to 'roll back' to a previous state. Rollback may be a case of reverting to a previous version of the object (e.g., in the case of files) or of 'undoing' a series of edits (e.g. in the case database with granular versioning). In the case of data at extreme scales it may not be possible to retain all deposited data or to work only on a copy. The version model must clarify any 'destructive' curation and the level of 'rollback' possible to previous versions.

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

- Published releases to end users
- Data products from linked data sources as independent, derived objects which require versioning (ideally with lineage of source objects)
- Versioning of complex collections of objects e.g. a series of longitudinal studies and its related documentation artefacts
- Dependencies on other objects such as funding, projects or authors
- Dependencies on other systems such as controlled vocabularies or file format registries

The minimal expectation is to:

- describe the operations that have invoked a version change of an object (from state A to state B)
- Describe the changes that have occurred between object in state A and object in state B
- Describe the version relationship between a source and a derived object (HasPreviousVersion, HasNextVersion)

## Negotiation/Acquisition

The negotiation process over which objects are accepted into a repository system clarifies which past versions and version metadata are offered for deposit. It should be clear to both parties which versions/metadata will be accepted and which version protocols will be continued by the repository.

A clear version/edition or wave model may be agreed for end users. A survey developed in an editor could be deposited with the questions' editing version history intact, but this version model may not be maintained if the repository transferred the data into a graph database for processing, which might not be able to mirror the structure and content of the versioning information. Upon export from Google docs to Microsoft word for deposit the version information is lost. Whether the depositor and repository choose to source the revision history via the Google Drive API[61] depends on the perceived value of provenance information at this level of granularity.

If the deposit references other objects (PID, author IDs, related publications) or has dependencies on other systems (software, XML schema, controlled vocabularies) that influence future versioning of the deposit, this should be clarified.

The level of curation and whether the repository is accepting long term preservation responsibility impacts future versioning decisions.

- A. Content distributed as deposited
- B. Basic curation – e.g., brief checking, addition of basic metadata or documentation
- C. Enhanced curation – e.g., conversion to new formats, enhancement of documentation
- D. Data-level curation – as in C above, but with additional editing of deposited data for accuracy

Whether basic curation (B) impacts versions depends on whether the additional metadata and documentation are treated as part of the object. Preservation responsibility (C-D) implies more complex versions while any "editing of deposited data for accuracy" (as agreed with the depositor/rights holder) implies a need to clearly document amendments at the data

[61] https://developers.google.com/drive/api/v3/manage-revisions?hl=en

point level”

Object characteristics which imply future version changes should be clarified. This includes at-risk formats not suitable for preservation or dissemination, but also covers deposits containing personal data where original and anonymised versions may be shared. Clear rights and IPR information defines whether these versions are accessible under different conditions (from Open Access, to secure remote access or Safe Room). Clear (and ideally machine actionable) rights and IPR information is critical if the repository is to mediate data linkage. The application of a rights modelling standard such as ODRL may support coherent and consistent permission management (L’Hours et al., 2018)(L’Hours et al, 2018b)

## Deposit and Curation

Sample data may be deposited and evaluated. Each deposit process may be handled through the creation of an ‘acquisition object’ by the repository with each version change being either:

- A new object based on the previous one
- An update to an existing object and storing the differences (commonly referred to as “diffs” or “deltas”)

Updates may be:

- Additions
- Changes
- Destruction or deprecation of some or all of object

The outcomes of each action should be defined and the level of detailed description of changes be agreed.

It should be clear to the repository whether they are amending (copies of) deposited materials or enriching them with additional materials. Enrichments are likely to be the intellectual property (IP) of the repository while amendments may cause IP to be shared between the depository and the repository. Changes to IP should be recorded in rights metadata.

Any changes which generate multiple related objects and data flows, such as the creation of anonymised versions of datasets, or the parallel processing of data in file, SQL and RDF environments should be recorded and the level of alignment between these paths over time defined.

Changes to anonymised data will also change the information security risk and access criteria for the new version. These should be recorded in rights metadata.

Co-dependent objects (PID, author IDs, related publications) or systems (software, XML schema, controlled vocabularies) that influence future versioning should be defined.

Define whether the object should be versioned independently or in conjunction with other objects or systems:

- Hierarchically: where a change to a child object updates the metadata and/or version of a parent object – “propagated versioning”
- Associatively: based on a change to a related object, such as adding a new related publication to an objects’ metadata.

Define the systems, such as controlled vocabularies of other registries of reference

information that the object depends on. Consider the versioning of those external systems and the local adoption of changes (see references to early and late binding in Versioning of surveys: the ESS Example above).

Consider the human and machine-mediated versioning approaches in place.

Define the agency (human actors or machine agents) permitted to undertake operations that invoke version changes.

Define which objects and parts of objects have associated identifiers and when a new version is required to have a new identifier.

Define which objects and parts of objects have change logs attached and what information must be recorded in the change log.

Identify which of these versions and associated version metadata need to be retained for internal repository use and which need to be exposed to end users.

Define the impact of version changes and how they should be communicated via internal systems and to end users (see below).

### Access and Use

At the point of access and use it is critical to communicate changes between versions to users of the data.

- Upon publication, relevant version, identifier and change metadata must be communicated to end users of the data.
- Only versions relevant to users will be shared (i.e. not intermediate curation versions or archival information package (AIP versions).
- Different levels of change impact must be clear through:
- Application and renewal of persistent identifiers
- A consistent numbering scheme
- Human-readable change logs
- Association with previous or subsequent versions

Local interpretations of high and low impact changes may vary, but a high impact change is likely to impact research outcomes and receive a new persistent identifier.

## Conclusions

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

The problem statement for repositories is to address version modelling for all representations of an object across the data lifecycle including consideration of interactions with co-dependant objects, and co-dependent systems. The version information must be presented without overwhelming the end user (depositor, curator or researcher) with redundant information that is not relevant to their needs.

Understanding the wider stakeholder ecosystem of actors and the local organisation context

supports good version model design. Designs should be applicable to a range of deposited data types so that human and machine processes can efficiently address a class of objects in a clear and consistent way.

The boundaries and version models for digital objects can vary with different perspectives so it is important for the deposit, curation and user actors to have a common vision of the objects to be managed e.g. multiple data collection waves could be treated as versions of a single study or a group of related digital objects. The chosen configuration will impact the versioning approach. The repository focus is to serve and align with the expectations of depositors and end users consistently.

Administrative and Social media data may include less version/provenance history at the point of deposit. For research data, including surveys, ideally at least the higher level versioning is defined in a data management plan and applied throughout study design and data creation/collection events.

While datastores of linked open data may have built in functionality to support very granular change logging or timestamping at the data point level this doesn't solve all the versioning challenges. As with verbose transaction logs offered as standard with classical relational databases, there is a usability challenge and data stewards must decide which version information (i.e. which digital objects in which states) are important and meaningful to each of their audiences.

Established information technology and research data management versioning practices are robust and will be applicable to the majority of challenges. But linked open data and the related technologies are in the early stages of developing common version management approaches and it will take time and research to identify how these versioned data products should be best presented to end users.

For a traditional file-driven archive where each 'study object' is commonly understood by both depositors and users (Wave X of Survey Y) and is managed through a single dataflow the presentation of new versions of objects with new identifiers and information about changes is relatively simple. In this scenario the internal versioning models of either the data depositor or the archive may be bespoke, local and non-transparent. As the number of partners who have data custody at some point in the lifecycle increase, or as a single repository manages instances of a dataset in multiple formats/environments (as file objects, SQL database or linked open data) the need to communicate and align version approaches increases.

Clear version models will also be a feature of transparency about processing undertaken and of clarity on the 'technical and organisational measures' which must be taken to protect personal data under GDPR. The maintenance of common information about each instance of a 'dataset' including its location and state will also be necessary for those seeking to certify their processes as trustworthy and their data as FAIR.

With Trusted Digital Repositories supporting findable, accessible, interoperable and re-usable data as part of the rules of engagement with the European Open Science Cloud there will be a need for cooperating peers to share data management techniques with each other. This should include clear, formalised approaches to object and version models with a view to alignment where this supports collaborative services.

# Appendices

# Appendices: EOSC, FAIR data, Preservation and the GDPR

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

## Appendix A: Stakeholder Ecosystem- Snapshot

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

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

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

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

### The EOSC Declaration & Implementation Roadmap

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

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

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

The declaration envisages an implementation roadmap supporting the move towards:

- Clear governance framework
- Definition of the initial services

- Clear business model
- Cost optimisation

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

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

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

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

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

The envisaged services are:

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

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


**Figure 2: EOSC Model Action Lines**

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

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

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

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

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

### Recommendation on Access and Preservation

See Appendix B for additional detail on this recommendation.

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

research outputs from all phases of the research life cycle (data, publications, software, methods, protocols, etc.)". Recital 7 states:

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

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

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

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

## Prompting an EOSC in Practice

The interim report and recommendations from the High Level Expert Group (HLEG) on the EOSC[62] demonstrates the current challenges for repositories in seeking to standardise their approach to handling NNfD. Released in mid-June 2018 this interim report indicates that the final version will define the features of an EOSC 'minimum viable ecosystem' (p20) and outlines initial thinking on Rules of Participation for federating existing infrastructures (p28) including consideration of private sector users stating "By participating, private sector may want to invest in the long term development and sustainability of the EOSC, along with the public sector and not just serve to exploit public data for free".

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

62 https://ec.europa.eu/info/sites/info/files/conferences/eosc_summit_2018/prompting_an_eosc_in_practice_eosc_hleg_interim_report.pdf

minimisation and protection of subjects' rights)

### Turning Fair Data into Reality

The integration of the FAIR data requirements into organisational and digital object management is presented in the interim report from the European Commission Expert Group on FAIR data: Turning FAIR data into reality (Hodson et al, 2018).[63] The report makes a clear early statement that the concepts of Findable, Accessible, Interoperable and Re-Usable are an excellent distillation of the mission and goals, but do not provide the best top-tier from which to consider implementation. A "holistic and systemic approach and to describe the broader range of changes required to achieve FAIR data" (ibid, p3) is taken. The report consists of 31 Metrics with Rec 1-14 incorporated into the executive summary. It has not been possible to fully analyse the implications due the recency of release, but this report and its successor will impact repository approaches to workflows and digital object management, including approaches to NNfD.

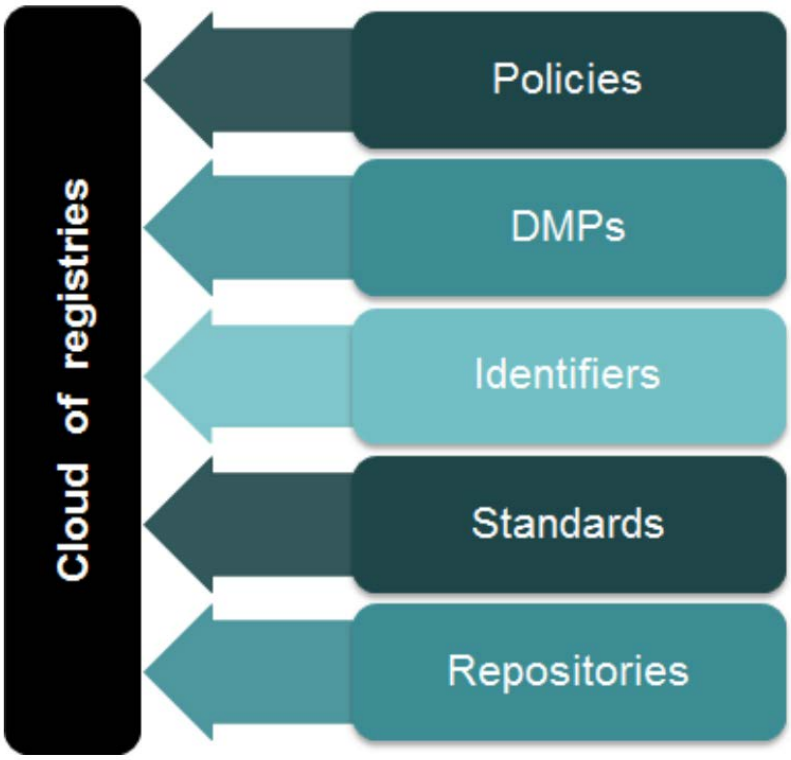


**Figure 3: Components of a FAIR data ecosystem**

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

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

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

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

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

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

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

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

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

This section is a revised version of a post published in cooperation with the Digital Preservation Coalition (DPC)[64].

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

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

### Mechanisms and Reporting

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

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

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

**Recitals**

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

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

(7) also notes that this has *traditionally* been the remit of archives and libraries, though we would expect that the vision of the European Open Science Cloud (EOSC) as a "trusted, open environment for the scientific community for storing, sharing and re-using scientific data and results" is seen as a common infrastructure component rather than a replacement for the 'traditional' centres of preservation practice... The quoted statement from the European Cloud Initiative states that this will "start by federating existing scientific data infrastructures, today scattered across disciplines and Member States" (8).

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

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

**Recommendations**

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

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

*Open access to scientific publications*

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

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

*Management of research data, including open access*

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

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

*Preservation and re-use of scientific information*

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

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

*Infrastructures for open science*

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

*Skills and competences*

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

*Incentives and rewards*

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

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

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

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

### Data culture and FAIR data

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

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

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

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

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

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

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

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

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

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

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

### Research data services and architecture

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

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

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

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

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

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

### Governance and funding

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

## Appendix D: Data, Processes and the GDPR

### Context

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

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

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

The official legislation page is at:

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

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

## Introduction

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

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

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

## The Digital Objects

> 1) 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;
>
> http://www.privacy-regulation.eu/en/article-4-definitions-GDPR.htm

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

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

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

**Contain personal data**

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

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

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

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

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

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

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

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

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

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

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

Pseudonymised data remains personal data under the GDPR.

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

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

> "To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and

> the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.
>
> The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."
>
> http://www.privacy-regulation.eu/en/recital-26-GDPR.htm

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

### Workflow Artefacts

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

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

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

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

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

### Actors and Roles

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## Context for Actions

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

'**lawfulness, fairness and transparency'**

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

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

Integrity and Confidentiality, *data security*

Demonstrable compliance for accountability (maintenance of records)

### Derogations from subjects' rights

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

**Research and archiving derogations** from the subjects' rights to:

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

16. **Rectification**

18. **Restrict Processing**

21. **Object**

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

**Additional archiving derogations** from the data subjects' rights to:

20. **Portability** of data

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

# Bibliography

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

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

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

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

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

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

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

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

# References


Allsop, S., & Shipsey, P. (2007). Version identification: a literature review, 37.

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

DASISH, Prestage, Y., Harrison, E., & Hq, E. (2015, January). Towards a common metadata understanding for the three DASISH WP3.2 tools.

DataCite Metadata Working Group. (2017). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data v4.1. https://doi.org/10.5438/0014

DDI Alliance. (2014, February 15). Data Documentation Initiative (DDI) Technical Specification Part I: Technical Documentation Version 3.2.

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

Kovalick, A. (2012). A Quick Tour of Wrappers and MXF, 2.

Krejčí, J. (2014). *Introduction to the management of social survey data*. Prague: Institute of Sociology CAS.

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

L'Hours, H., Butt, S., Henriksen, G., Krejčí, J., Štebe, J., Myhren, M., … Bell, D. (2018b, August). Appraisal/Selection Requirements for New Forms of Data. Deliverable 6.9 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme GA No: 654221.

OECD. (2016). *Research Ethics and New Forms of Data for Social and Economic Research* (OECD Science, Technology and Industry Policy Papers No. 34). https://doi.org/10.1787/5jln7vnpxs32-en

PREMIS Data Dictionary for Preservation Metadata, Version 3.0. (2015). PREMIS Editorial Committee.

Rauber, A., & Asmi, A. (2016). Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use, 10.

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

Versions Toolkit for authors, researchers and repository staff. (2008, February). LSE. Retrieved from http://eprints.lse.ac.uk/64948/1/LSE%20Library_Versions%20toolkit_Author_2008.pdf