Research Data Management Guidance Manual

Milan Slavkovic

July 2024

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

CONTENTS

1. STATEMENT OF PURPOSE AND MOTIVATION

In the past decade, the manipulation of digital data resources has significantly grown in importance across various research fields, leading to a crucial need for proper planning regarding resource allocation and budgeting for data management. Science funders and research institutes now require data management plans (DMPs). Funders demand DMPs to ensure that data collection expenditures benefit other researchers, while research institutes seek them to uphold scientific integrity and reproducibility standards. However, many researchers view DMPs as burdensome obligations rather than effective tools, due to limited awareness of available tools and expertise for data management and a lack of experience in recognizing data-related risks. Additionally, researchers' experiences with digital data in personal settings often fail to adequately prepare them for managing the complexities and collaborations involved in research data management, while experts providing data management services struggle to connect with researchers in need of their expertise.

After surveying leading research institutions in Serbia, the findings showed that various organizations employ diverse strategies for data management and security within their respective institutions. As they fall under the University of Belgrade's umbrella, they share certain commonalities, such as institutional repositories regulated by the University of Belgrade's Computer Centre - RCUB. These repositories serve as a strong foundation for adhering to open science policies and FAIR (findable, accessible, interoperable, reusable) principles for data. Moreover, initiatives like SAIGE aim to develop DMPs for the majority of institutions, providing a robust framework for raising awareness among scientists. Certain organizations benefit from partnerships with established international entities through various international projects and collaborations where a DMP is required to apply for grants under any grantee R&D and R&I entity project (e.g., the Horizon Europe program). While there are some similarities, the extent of data management practices and the strategies employed to safeguard data differ greatly among institutions. Therefore, institutions within the University of Belgrade could derive significant advantages from adopting standardized best practices and procedures for data management, as outlined in this manual.

This guidance manual should serve as a robust foundation for the creation of a data management plan (DMP) for all research projects at the institutional level. Also, covering essential aspects of data management and security, the manual addresses secure methods for collecting, storing, sharing, transferring, and disposing of research data. As defined by OECD/G7: research security is safeguarding the research enterprise against the misappropriation of research and development to the detriment of national or economic security, related violations of research integrity and foreign government interference. This guidance outlines measures for safeguarding research data against unauthorized access, use, disclosure, disruption, modification, or destruction while providing an overview of Research Data Management (RDM) practices within research organizations. It is vital research institutions incorporate secure practices for research data to avoid the loss of institutional assets through forced technology transfer or intellectual property theft, as research data serve as the foundation of the research enterprise.

1.1. Executive Summary

The Research Data Management Guidance Manual (RDMGM) is structured according to relevant standards such as Horizon Europe, ISO/IEC 27001, NIST Special Publication 800-171, and FIPS 140-2, along with recommendations from organizations and working groups established by the European Commission for RDM and OA, as documented in the literature review. The manual is thematically divided according to the data lifecycle, which follows the course of the research process in three parts: before, during, and after research. The starting chapter (2) covers the planning phase

of data management. In the following (3) chapter, which focuses on data management during research, best practices for working with data in the phases of collection, processing, and analysis are presented. The final part of the manual, "After Research," provides an overview of the most important recommendations related to permanent storage, sharing, and utilization of research data. To facilitate effective and sustainable data management in collaborative projects, the RDM-GM provides appendices that serve as a guide for developing DMP, offering recommendations and tools for researchers, and a comprehensive resource for data security and management in organizations. Instructions for answering the questions are highlighted in yellow. The list of appendices includes:

- Appendix A – Project datasets template
- Appendix B - Guideline questions for developing the data summary section of DMP
- Appendix C – Guideline questions for developing the data fair verification section of DMP
- Appendix D – Guideline questions for developing the other research outputs section of the DMP
- Appendix E – Guideline questions for developing the allocation of resources section of the DMP
- Appendix F– Guideline questions for developing the data security and privacy section of the DMP
- Appendix G - Guideline questions for developing the ethics issue section of the DMP
- Appendix H – Recommended and selected repositories and databases
- Appendix I - Sample metric questions for assessing security awareness and training efforts
- Appendix J – Informed consent form example

1.2. Description of Research Data and Data Lifecycles

Research data includes all data collected, recorded, or generated to produce new, original research results. In that sense, research data can be classified[1] as:

- Raw (initially collected)
- Cleaned (prepared for analysis)
- Processed (data resulting from conducted analysis)
- Presentation (data adapted for presentation).

The research ecosystem is becoming increasingly complex, leading to a growing need for improved collaboration within specific research communities, both interdisciplinary and involving the public - citizen science. All of this underscores the necessity for sharing research data, which should be addressed not only at the end but also throughout the entire research process. Benefits of sharing research data include:

- Contributes to the reproducibility and transparency of scientific research.
- Increases the impact of scientific research and citation rates.
- Opens up possibilities for collaboration on other research projects.
- Meets the requirements of research funders.
- Enables their use for educational purposes in teaching students.

[1] University of Edinburgh, Research Data Service: Our definitions. URL: https://www.ed.ac.uk/information-services/research-support/research-data-service/after/data-repository/definitions (2020-08-28)

Figure 1: Data Lifecycle Model (UK Data Archive)

Figure 1 depicts a model of the data lifecycle in research. To ensure that data remains understandable and reusable both during and after research, it is necessary to manage them throughout all phases of the lifecycle. It typically includes stages such as data collection, where information is gathered from various sources. Next comes data processing and analysis, where the information is manipulated and interpreted to derive insights. Subsequently, data may be disseminated to relevant stakeholders for decision-making purposes, followed by storage, where data is stored securely in appropriate repositories. Finally, the data is either archived for future reference or deleted in accordance with data retention policies and regulatory requirements. This model ensures that data is effectively managed, protected, and utilized throughout its lifecycle, maximizing its value while minimizing risks associated with data breaches or misuse.

Also, Figure 2 illustrates how the data migration process combines human and computer actions. Treloar and Harboe-Ree (2008) note that researchers generally do not prioritize data curation, a task best handled by professionals in the publication domain. Successful public data curation depends on provenance metadata collected during research. However, the necessary skills and tools for effective research data management are not well-defined. Both the data life cycle and curation frameworks emphasize sharing, suggesting that digital curation practices are more prescriptive than descriptive.

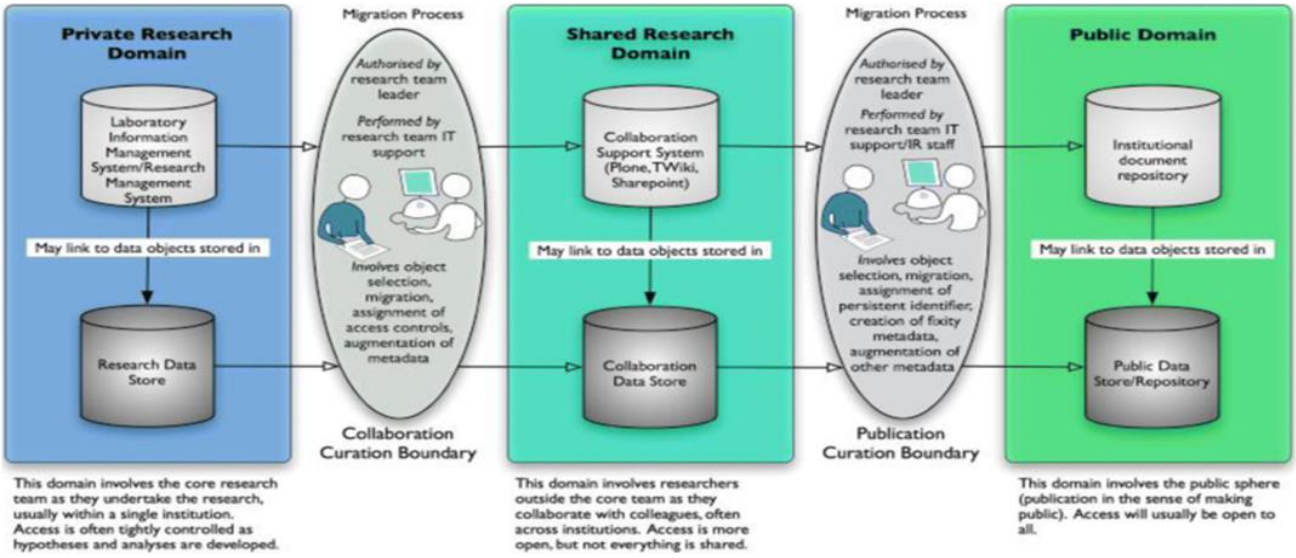

Figure 2: Data curation continua. In Treloar and Harboe-Ree 2008, pg.6

2. BEFORE BEGINNING RESEARCH (LIFECYCLE: PLANNING)

Before initiating research, it is crucial to strategize the management of research data, known as Research Data Management[2] (RDM). RDM involves handling data generated throughout the research cycle and forms an essential component of the research process, assisting researchers in organizing, describing, storing, and sharing data effectively.

As part of research planning, it's imperative to include data management planning, which is significantly aided by the development of a document known as the Data Management Plan (DMP). This is the most important step in the data lifecycle to incorporate management and security best practices as this will dictate important aspects of the overall research project.

The goal is to establish and routinely update a Data Management Plan[3] (DMP). This plan not only aids in structuring research data but also promotes quality and efficiency in data management, potentially resulting in time and resource savings in the long term. The DMP outlines protocols for securing data, including encryption, access controls, and backup procedures. It ensures compliance with regulations, promotes accountability, and fosters a culture of awareness regarding data security among stakeholders. Furthermore, it assists in the preservation and dissemination of research results data, determining which research will be shared, when, and how to manage sensitive or commercially valuable information. Sharing research findings can enhance productivity through increased citations and public recognition of research endeavors. Additionally, a shared dataset facilitates collaboration among researchers, expediting discoveries in intricate research domains.

The purpose of the DMP extends to grant applications, such as those within the Horizon Europe program, where it is mandatory as a supplementary document. Its primary objective is to ensure the availability and usefulness of institutional project research data for future purposes and applications, tailoring the plan to suit the specific characteristics of each project.

The scope and limitations of the DMP involve outlining the types of data and research outputs in organizational projects, ensuring compliance with FAIR data principles (Findable, Accessible, Interoperable, Reusable), and specifying data storage and preservation approaches. Additionally, the plan addresses adherence to Open Access Policy requirements for disseminating and sharing R&D results.

The subsequent sections outline guidelines that researchers should incorporate when formulating a DMP, based on recommendations from the Horizon Europe program's Data Management Plan.

2.1. Section 1 – Data Summary

The DMP[4] should provide a concise overview of the data to be collected and generated throughout the project, including their types, formats, approximate scope, any sensitivities, and potential sources of secondary data from open repositories, databases, patent databases, etc.

Moreover, during experimental research, especially in the stages of methodological design and research data analysis, there is a focus on gathering and analyzing open research data linked to published papers in scientific journals[5] (accessible through available data sections or supplementary materials), along with data stored in general or domain-specific repositories. Researchers should

[2] Research Data Management Definition available at: https://libguides.depaul.edu/c.php?g=620925&p=4324498
[3] OpenAIRE guidance: https://www.openaire.eu/how-to-create-a-data-management-plan
[4] https://www.scienceeurope.org/media/4brkxxe5/se_rdm_practical_guide_extended_final.pdf
[5] https://open-research-europe.ec.europa.eu

adhere to the terms of the reuse license and attribute the original data to its creator(s) appropriately. Initially, the intention is to explore deposited data within the repositories (Please see: Appendix H).

Given the increasing significance of open data within the framework of Open Science policies, there is a planned effort to search for and utilize existing datasets that are specifically curated for this purpose. The search will involve utilizing one of the following methods, portals, or providers:

- DataSearch :
  Using a simple keyword search, users can discover datasets hosted in thousands of repositories across the Web.
- DCU library :
  Collated number of dataset sources.
- DataEurope:
  Provides access to 1.4 million public datasets from 36 countries (EU, EEA, Switzerland and EU Neighbourhood states). Data resources are indexed by the European Commission from national, regional, local and domain-specific public data providers. The interface is available in six languages.

Typical types and formats of data handled in research projects include:

> **Types of data:** qualitative and quantitative data from experimental measurements, as raw, abstracted, or analyzed data; numerical data (datasets, spreadsheets); textual data (protocols, methodological descriptions, laboratory notebooks, field notebooks, documents, reports); programming data (web-based data), spectral data and DNA sequences, mixed media data (image, audio, video); meta-data; and presentations.
> **Formats of data:** .dat, .txt, .opj, .xlsx, .csv, .jpg, .png, .bmp, .spc, .fasta, .cmb, .docx, .pdf, .pptx, .html, .avi, .mp4, flac, .zip, .json, PDF/A etc.

Researchers should be committed to aligning the selection of data formats with the principles of good open science practice and specific standards endorsed by data repositories and the research community. This approach aims to enhance the sharing and long-term reusability of data under FAIR[6] principles. While certain data formats may be influenced by the requirements of particular software or laboratory equipment, detailed guidance for such formats should be included in the readme.txt file. This guidance should encompass information such as the software name, license version, potential download link, and installation instructions, ensuring accessibility and usability for interested parties.

Numerous repositories offer curated lists of preferred formats designed to facilitate successful long-term preservation and reusability. Consequently, a broader discussion on this topic is provided in Section 3.1.4. as Table 3 outlines the recommended formats tailored to the common types of data expected to be generated.

An additional consideration for the Data Summary is the identification of any sensitive data or restrictions on information release. Including this early in the project is important to maintaining security practices in line with the risk of data loss or misuse. Examples include Personally Identifiable Information, classified or restricted data, or data under the purview of export controls.

As the DMP evolves as a living document, the data summary should be periodically adjusted in subsequent versions to reflect the digital content generated throughout the research or project. The

[6] https://fairsharing.org/

recommended version of the DMP includes Appendix A, which presents a preliminary list of datasets to be generated, along with concise descriptions and associated open-access conditions. Additionally, Appendix B comprises questions aimed at aiding the development of the DMP concerning data summary. Similar annexes containing questions relevant to specific sections will be included in subsequent sections of the DMP.

2.2. Section 2 – FAIR Data

This section elaborates on the implementation of the FAIR[7] principle in detail, outlining procedures and approaches as far as possible at the project's outset. These should be refined based on feedback received during the specified period of research or the project.

2.2.1. Making Data Findable, Including Provisions for Metadata

All research data representing digital content in various formats is recommended to undergo a review by the designated project boards, e.g. Executive Board (EB), to determine their suitability for public access. Data lacking any restrictions, such as those related to data protection, privacy, confidentiality, trade secrets, Union competitive interests, security, or intellectual property rights, should be accompanied by comprehensive metadata aligned with FAIR principles (please see Appendix C). The fundamental approach is to assign a persistent identifier (PID)[8], such as a Digital Object Identifier (DoI) to each dataset. Persistent Identifiers (PIDs) play a crucial role in data security by providing a unique and persistent reference to digital objects, ensuring their integrity and traceability across various systems and platforms. PIDs are also a valuable tool in the event data is stolen or manipulated to identify what has been compromised and how to direct recovery efforts. To facilitate this, trusted repositories offering automatic PID assignment should be selected. Some of the recommended repositories are listed in Appendix H.

Given the anticipated volume of research data to be generated throughout the project, internal protocols should be implemented for organizing the data. These protocols will encompass procedures for distributing data among files, standardizing file naming conventions, and organizing directories. Further exploration of this topic is available in Section 3.1, where a comprehensive discussion and guidelines for organizing research data are presented.

2.2.2. Making Data Accessible

To ensure open access[9] to publications, beneficiaries must deposit them into a repository and choose between 'gold' OA publishing or 'green' OA self-archiving. Gold OA involves immediate publication in open-access journals or hybrid journals, with publication costs covered by authors. Green OA allows self-archiving of peer-reviewed manuscripts into repositories, potentially subject to embargo periods. Additionally, providing access to bibliographic metadata alongside publications is crucial for facilitating data reuse. Access rights to background data, essential for project implementation or result exploitation, must be agreed upon by beneficiaries, with options for fair and reasonable conditions. Trans-national access to research infrastructure typically includes royalty-free access to background data, with access rights varying based on project needs and agreements between parties. EU institutions and Member States typically do not possess distinct access rights, except for royalty-free options, with access typically limited to beneficiaries and affiliated entities. A

[7] https://www.go-fair.org/fair-principles/
[8] https://open-research-europe.ec.europa.eu/for-authors/data-guidelines#persistentidentifier
[9] http://opendatahandbook.org/guide/en/appendices/file-formats/#open-file-formats

more in-depth examination of this subject can be found in Section 4.1, which offers a thorough discussion and provides guidelines for aligning with the open-access policy of research data.

2.2.3. Making Data Interoperable

Data Objects can be Interoperable only if:

- (Meta)data is machine-actionable[10];
- (Meta)data formats utilize shared vocabularies and/or ontologies; [11]
- (Meta)data within the Data Object should thus be both syntactically pursuable and semantically machine-accessible[12]

In order to make research data interoperable, data produced within the project should be formatted and stored using standard formats widely recognized in the scientific community. These formats should facilitate seamless integration into existing software applications for subsequent utilization. Standardizing the vocabulary and methodologies employed across the project ensures interoperability, enhancing efficiency and collaboration. Regarding the datasets derived from experimental research, predominantly comprising well-organized numerical data in tabular form, they should initially be stored internally in the XLSX format. However, to deposit them in open repositories, they should be transformed into machine-readable CSV (Comma Separated Value) format, which is also actionable. This is only possible if each spreadsheet file is prepared appropriately. In this regard, the following approach should be applied:

- Each column should have a descriptive heading in single raw
- Each worksheet should start from the first cell A1
- Each spreadsheet should have a title and a legend
- Each worksheet is a separate file to be deposited within the dataset
- Non-alphanumeric characters, including commas, should not be used
- Spreadsheet should not have merged cells and colour coding
- Any chart, comment, or other tables should be excluded from the spreadsheet

Additional research data, not of numerical nature, will also be formatted into machine-readable formats, aligning with established standards such as RDF, XML, or JSON. Many repositories offer assistance in verifying and formatting data according to these standards; thus, this factor will also influence the selection of a trusted repository.

2.2.4. Increase Data Re-use

For Data Objects to be Re-usable, additional criteria include:

[10] https://force11.org/info/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-0/#Annex6-9

[11] https://force11.org/info/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-0/#Annex6-9

[12] https://force11.org/info/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-0/#Annex6-9

- Data Objects should be compliant with principles 1-3 [13]
- (Meta)data should be sufficiently well-described and rich that it can be automatically (or with minimal human effort) linked or integrated, like-with-like, with other data sources [11 and Joint Data Citation Principles: JDDCP7- and JDDCP 8][14]
- Published Data Objects should refer to their sources with rich enough metadata and provenance to enable proper citation[15]

The structure of dataset metadata should include additional information related to research methodology, sampling, experimental conditions, variables, and units of measurement, as well as procedures for processing and analyzing experimental data, which enables validation of the conducted analysis and displayed results, as well as re-use of the data. Annex II comprises questions aimed at aiding the development of the DMP concerning FAIR principles.

Important to mention here that anytime datasets or meta-data will be reused it is also relevant to assess if this re-use will result in dual-use research.

2.3. Section 3 – Other Research Outputs

Increasing transparency in the research process and its outcomes is key to receiving more citations, which is important for both researchers and the reputation of companies. However, this largely depends on how complete the research output is, how well it's documented, where it's stored, and how it's shared[16]. In the project, besides generating data from experiments, there will be significant research findings. So, it's crucial to plan how to share these findings according to the requirements of the Grant agreement. Usually, sharing these findings might happen months or years after the data is collected. Some processes, like getting a paper published in top scientific journals or getting a patent approved, take time. Hence, the DMP may only offer a preliminary plan. It's hard to predict when innovations will be completed and protected for sharing through patent databases or other means, as research projects involve uncertainties and risks. This section considers the project's innovative potential and associated constraints. This section is closely linked with the Dissemination, Exploitation, and Communication Plan and should be updated in subsequent versions of the Plan.

2.4. Section 4 – Allocation of Resources

The adoption of new open science practices in Higher Education (HE), including the Data Management Plan (DMP), requires significant resources for successful implementation in terms of human, financial, and material resources. Managing research data and outputs is integral to the research process throughout the project, containing various aspects such as data collection or acquisition, curation, storage, preservation, data security, quality assurance, allocation of persistent identifiers (PIDs), metadata provision, licensing, and establishing rules and procedures for data sharing. Particularly significant efforts are required for data management in research work packages where substantial amounts of data are collected and generated.

[13] https://force11.org/info/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-0/#Principles1-3
[14] https://force11.org/info/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-0/#Principles1-3
[15] https://force11.org/info/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-0/#Principles1-3
[16] https://onderzoektips.ugent.be/en/tips/00002153/

Consequently, a portion of the staff costs outlined in the project budget should be allocated to these activities.

2.5. Section 5 – Data Security

Data management and security are most effective when they are incorporated by design rather than applied during the course of research or as an afterthought. Throughout the research project, it is imperative for all parties involved to maintain confidentiality regarding any data, documents, or materials identified as confidential upon disclosure. Beneficiaries have the option to extend the confidentiality period and agree upon additional confidentiality-related obligations as needed. Data security includes various dimensions beyond confidentiality, extending to aspects such as Dual Use Research of Concern (DURC) and export controls, sensitive research data (including genomic data or Personally Identifiable Information (PII) from surveys), and even military-contracted research. In these contexts, stringent measures are necessary to safeguard data integrity, prevent unauthorized access, and mitigate the risk of misuse or exploitation. For DURC and export-controlled data, compliance with regulatory frameworks and international agreements is essential to prevent the dissemination of sensitive information that could pose risks to national security or public safety. Additionally, in sensitive research areas such as genomics or surveys involving PII, data security measures must be robust to protect individuals' privacy rights and prevent potential harm resulting from unauthorized disclosure or misuse of sensitive information.

To safeguard research data[17], including raw, processed, and analyzed data sets, it is advisable to securely store them within decentralized institutional storage systems managed by their respective IT teams. These systems ought to support data saving, backup, and sharing among project collaborators while discouraging storage on personal laptops, computer hard drives, or external devices, even if encrypted. Researchers are mandated to routinely transfer gathered and processed research data to institutional storage, with authorized access managed by IT teams. Reliable storage systems equipped with automated backup features guarantee data permanence and facilitate recovery in the event of unforeseen incidents, with partners adhering strictly to institutional data protection protocols. In cases where researchers from different partner institutions collaborate on the same task, data may be temporarily transferred via secure FTP protocol or utilized on recommended cloud-based storage solutions such as Amazon Web Services, Microsoft Azure, or Google Suite for Education. It is emphasized that the use of certain platforms like Dropbox, personal Google accounts, personal OneDrive accounts, or Apple iCloud is not advisable.

Furthermore, long-term preservation of analyzed and final data will be facilitated by project partners, either internally or through external trusted repositories where the data is deposited (Section 3.2.4).

Also, consider what type of access requirements will be required for your research data based on the personnel involved and the nature of the project (Section 3.2.3).

2.6. Section 6 – Ethics

The projects should be designed to meet both ethical and legal standards, ensuring that their outcomes are embraced and utilized by end-users.

Ethical considerations are paramount throughout the project's implementation. The project must adhere strictly to the guidelines outlined in the Grant Agreement, as well as relevant international and national laws, upholding the highest levels of research integrity and information security. A key

[17] https://cordis.europa.eu/project/id/101046758/results

ethical concern addressed is compliance with The General Data Protection Regulation[18] (GDPR). Personal data should only be collected when absolutely necessary for the advancement of the project, with informed consent obtained for data sharing and long-term preservation. Access to this information is restricted solely to team members, ensuring confidentiality and non-disclosure to third parties. Sensitive data should be securely stored.

2.7. List of Available Tools

- **Argos**— online tool for creating DMPs
- **Data Management Skillbuilding Hub** — database providing recommendations for working with data, covering every stage of the data lifecycle
- **DMPonline** — online tool for creating DMPs
- **Swiss National Science Foundation**, **European Commission**, **UK Data Service** - guidelines that researchers should consider when creating a DMP
- **Digital Curation Centre** , **Swiss Data Life-Cycle Management (DLCM)** - examples of DMPs

[18] https://gdpr-info.eu/

3. DURING THE RESEARCH (LIFECYCLE: COLLECTION, PROCESSING, ANALYSIS)

This chapter deals with the phases of the research lifecycle related to data collection, processing, analysis, and protection. Within these phases, special attention will be given to rules regarding the naming and organization of research data, controlling different versions of research data, preferred file formats for storing research data, documentation and metadata description creation, and data protection. As the data needed for research is collected, the number of files and folders containing research data increases. Finding the desired research data among numerous files stored on the local storage system or in the cloud can be challenging if the files are not consistently named or if there is no logical folder structure.

To make the process of managing research data more efficient, it is important to establish a naming convention for research data and an appropriate folder structure before the research begins. A naming convention for research data is a way of assigning names to files and folders related to research data. It is documented in the research data documentation, such as a readme file.

3.1. Recommendations Regarding the Labeling and Organizing of Research Data

3.1.1. Guidelines for Naming Research Data

- Develop a naming convention for research data consisting of logical elements. Relevant information for file naming includes project name or acronym, researcher's name or initials, data type, research method, location and date of research, and file version number.
- Consistently use naming conventions for research data.
- Utilize naming conventions adopted by all members of the research team.
- Avoid using similar names for multiple files.
- File names should not be overly long, not exceeding 32 characters.
- Use ASCII letters and numbers (a — z, A — Z, and 0 — 9). Avoid using special characters in file names, such as &, *, %, #, ;, ( ), !, @, $, ^, ~, ', { }, [ ], ?, < >.
- Use underscores, hyphens, or camel case instead of spaces.
- Use the ISO 8601 standard for dates (YYYYMMDD).
- Identify different file versions with numerical labels.
- Automated file renaming can be facilitated by software solutions, such as Bulk Rename Utility.

3.1.2. Guidelines for Organizing Research Data

- Establish a folder structure that meets the needs of the research project. It is important to consider the type of research data and how raw and analyzed data, methods, documentation, and other supporting files will be organized within it.
- Give unique names to folders that correspond to the research project.
- Achieve an appropriate depth of folder hierarchy. Excessively deep hierarchy will result in a large number of clicks to reach the desired file, while folders with overly shallow hierarchy will contain too many files.

- Tagging files helps in locating files within the research project folders. Tags are keywords assigned to files that should be simple and consistently used. Tagging and searching local files can be done, for example, within Windows and macOS operating systems or by using programs such as Adobe Acrobat or Adobe Bridge, and TagSpaces which is compatible with Windows, macOS, and Linux operating systems.[19]
- If working with collaborators, either internal or external to your organization, establish a protocol for version control of research project assets to ensure work is done on the appropriate data. This has the potential to save time and resources by avoiding confusion and miscommunication during a project.

Documentation is the process of describing research data, recording what the research data are, and how they were collected, analyzed, and versioned during the research. Since documentation provides context to research data, it is of great importance to both the public and the research team for understanding, sharing, and reusing them.

Metadata describes a set of research data (data about data). It is recommended to record metadata in a so-called readme file and update it during the research project. The readme file should be in a textual format (.txt), which is crucial for long-term protection and readability.

3.1.3. Guidelines for Metadata in the Readme File

To facilitate comprehension and utilization of the generated data by secondary users, the comprehensive metadata should be compiled in a "readme.txt" style for each dataset intended for open access. This metadata will serve as the foundation for publishing research findings. A suggested template for creating metadata, along with explanations, can be found here: READMEtemplate.

The dataset metadata will encompass essential details enabling their discovery by other users, whether computer or human. These details should include:

- **Descriptive Title:** Providing a descriptive title to help users identify the general content and purpose.
- **Persistent Identifier:** Issued by a selected reputable, long-term repository.
- **Creator(s) Names:** Including name, ORCID, institution, address, and email.
- **Geographic Location of Data Collection:** Indicating latitude, longitude, or city, region, and country.
- **Creation Date:** Using the format YYYY-MM-DD.
- **Spatial/Temporal Coverage.**
- **Types of Files in the Dataset.**
- **Metadata Standard Utilized** (if applicable).
- **Data Location**: Links to other publicly accessible locations of the data.
- **Funding Statement**: Details about funding sources supporting data collection, e.g., The data were generated as part of the research conducted within Project X, funded under the HORIZON EUROPE program (HORIZON-EIC-2023-PATHFINDEROPEN-01, GA no 12345678, Oct 2023 – Sep 2026.

[19] UK data services

- **Description – Data Abstract**: Succinct dataset description listing all files or folders, as appropriate for dataset organization, contained within the dataset, and any significant relationships between files.
- **Keywords.**
- **Data Licenses/Restrictions:** CC BY or CC0, considering constraints.
- **Related Publications:** Links to publications citing or using the data.
- **Sources** (if data is derived from other source(s)).
- **Recommended Dataset Citation:** Links to accessible locations of the data.

Additionally, any supplementary information and documentation enhancing dataset comprehension and facilitating its reuse may be included in metadata, providing details on:

- Methodology employed for data collection/generation
- Sampling procedures
- Variables, their definitions, and units of measurement
- Experimental conditions
- Instrument- or software-specific information necessary for data interpretation
- Standards and calibration information
- Processing and analytical steps conducted
- Quality-assurance procedures performed on the data.

It should be noted that specific disciplines and repositories may dictate the content and format of metadata. Considering the interdisciplinary nature of research within the project, it is necessary to define and incorporate additional descriptive information into dataset metadata throughout the project. OpenAIRE guidelines for data archiving can be consulted for this purpose.

3.1.4. Guidelines for Selecting Data Storage Formats for Long-Term Preservation

When storing research data, it is important to consider the formats in which the data will be stored and accessible. It is crucial to choose open and well-documented formats that are supported or can be supported by multiple software vendors, which facilitates potential migration to new formats in the future. Selecting an appropriate preferred or acceptable format ensures long-term data storage and availability. When choosing the appropriate type of format for data storage, it is recommended to use open and widely applicable formats such as plain text (ASCII, .txt), comma-separated values (.csv), or XML (Extensible Markup Language) instead of closed, proprietary formats such as Microsoft Office, Statistical Package for the Social Sciences (SPSS), and others. Table 3 provides several examples of open and closed formats.

Table 1: Overview of Data Storage Formats

Data Types	Digital Content	Recommended Formats	Other Acceptable Formats
Text	Protocols Documents Reports	Plain text (ASCII, UTF-8, UTF-16 with BOM) Open Office (.odt)	MS Office (.doc, .docx) PDF (.pdf)

Data Types	Digital Content	Recommended Formats	Other Acceptable Formats
		Reach text format (.rtf) PDF/A (Archival PDF)	
Numerical	Datasets Spreadsheets	ASCII or Unicode Comma Separated Values (*.csv) Delimited Text (*.txt) **Compressed format** ZIP	MS Office (*.xlsx, *.xls)
Experimental data	DNA sequences Spectral data	.fasta, .spc, .cmb	
Image	Raster image Micrographs	JPEG/JFIF (.jpg) PNG (.png) PDF/A, TIFF	JPEG/JPEG2000, BMP (.bmp)
Audio/Video	Audio/video recordings, moves etc.	FLAC, AIFF, WAVE M-JPEG2000 AVI (*.avi)	MPEG3 (.mp3) MPEG-4 (*.mp4)
Presentation	Power Point presentations		MS Office (.pptx, .ppts)

3.1.5. Best Practices for Selecting Formats for Long-Term Data Preservation

- Utilize open and well-documented formats.
- Opt for formats with support across various platforms (operating systems)
- Prefer textual formats over binary ones (.txt instead of .pdf).
- Avoid encryption or password protection of files when unnecessary, as it complicates their migration to other formats.
- Additionally, refrain from compressing files with lossy compression, a process that reduces the original file size at the expense of content quality.

When selecting a data storage location, it is necessary to consider available options and functionalities such as automatic backup, data sharing and exchange with others, and data encryption with partners adhering to institutional data protection policies, and data security best practices more broadly. Researchers should apply institutional data protection policies, and may also consider applying more robust data protection practices where gaps are identified and depending on the sensitivity of the project.

3.2. Guidelines for Data Protection Policies in Research Organizations

3.2.1. Data Collection Procedures

- Utilize specific guidelines for collecting data, including obtaining informed consent from participants and ensuring transparency about data collection practices. (Section 4.3)

- Outline procedures for anonymizing or de-identifying personal data to protect the privacy of research subjects. (Section 4.4)
- Define protocols[20] for securely transferring data from data sources to research repositories:
    - Secure File Transfer Protocol (SFTP):
        - SSH Communications Security - SFTP
    - HTTPS:
        - MDN Web Docs - HTTPS
    - Secure Shell (SSH):
        - OpenSSH Project
    - Aspera[21]:
        - IBM Aspera Documentation
    - VPN[22]:
        - OpenVPN Documentation

These sources should provide detailed information about each protocol and how to implement them for securely transferring data in research environments.

3.2.2. Data Storage and Retention Procedures

3.2.2.1. Define standards for storing research data securely, including encryption protocols and access controls.

Encryption protocols and access controls, as specified by ISO/IEC 27001 and NIST Special Publication 800-171, include:

3.2.2.1.1. ISO/IEC 27001

Encryption Protocols: ISO/IEC 27001 recommends the use of encryption to protect data in transit and at rest. Common encryption protocols include AES (Advanced Encryption Standard) for symmetric encryption and RSA (Rivest-Shamir-Adleman) for asymmetric encryption.

Access Controls: ISO/IEC 27001 emphasizes the implementation of access controls to ensure that only authorized individuals have access to sensitive information. This includes user authentication mechanisms such as passwords, biometrics, or multi-factor authentication, as well as role-based access control (RBAC) to restrict access based on user roles and responsibilities.

3.2.2.1.2. NIST Special Publication 800-171

Encryption Protocols: NIST SP 800-171 requires the use of encryption to protect CUI (Controlled Unclassified Information) stored in information systems and transmitted over networks. It specifies cryptographic standards and algorithms for encryption, including AES for symmetric encryption and RSA for asymmetric encryption.

[20] https://security-guidance.service.justice.gov.uk/secure-data-transfer-guide/#data-transfer-by-email
[21] https://www.applytosupply.digitalmarketplace.service.gov.uk/g-cloud/services/549451152565304
[22] https://www.forbes.com/advisor/business/software/vpn-protocols/

Access Controls: NIST SP 800-171 outlines access control requirements to limit access to CUI only to authorized users and processes. This includes user authentication, access permissions, access monitoring and logging, and session management controls.

3.2.2.2. Specify the types of storage devices and systems approved for storing research data, such as secure servers or cloud storage platforms.

3.2.2.2.1. Secure Servers

Secure servers refer to dedicated hardware devices or virtual machines designed to store and manage data securely within an organization's network infrastructure. Secure servers are often recommended by information security standards such as ISO/IEC 27001 and NIST SP 800-171 for storing sensitive research data.

Advantages:

- Centralized management: Secure servers allow for centralized management of data, making it easier to implement security policies and access controls.
- Customizable security configurations: Organizations have greater control over security configurations on dedicated servers, allowing them to tailor settings to their specific requirements.
- Enhanced performance: Secure servers typically offer higher performance compared to other storage solutions, especially when handling large volumes of data or high-intensity workloads.

Disadvantages:

- Cost: Setting up and maintaining secure servers can be expensive, requiring investments in hardware, software, and skilled personnel for administration.
- Physical infrastructure: Organizations need to ensure proper physical security measures for server rooms or data centers to prevent unauthorized access to the hardware.
- Maintenance complexity: Secure servers require regular maintenance and updates to address security vulnerabilities and ensure optimal performance, which can be time-consuming.

3.2.2.2.2. Cloud Storage Platforms

Cloud storage platforms offer scalable and convenient storage solutions hosted by third-party service providers. They allow organizations to store data off-site and access it over the internet.

Advantages:

- Scalability: Cloud storage platforms offer scalable storage options, allowing organizations to easily adjust storage capacity based on their needs without significant upfront investments.
- Accessibility: Data stored in the cloud can be accessed from anywhere with an internet connection, enabling remote collaboration and data sharing among research teams.

- Built-in security features: Many cloud providers offer built-in security features such as encryption, access controls, and compliance certifications to protect data against unauthorized access and data breaches.

Disadvantages:

- Data residency and compliance concerns: Organizations need to consider data residency requirements and regulatory compliance when storing sensitive research data in the cloud, especially if data sovereignty is a concern.
- Dependency on the service provider: Organizations rely on cloud service providers to maintain the security and availability of their data, which can pose risks if the provider experiences downtime or security incidents.
- Potential for vendor lock-in: Migrating data between cloud providers or back to on-premises infrastructure can be challenging, leading to vendor lock-in and limited flexibility.

3.2.2.2.3. Network Attached Storage (NAS)

NAS devices are specialized file servers connected to a network that provides centralized data storage and access to multiple users and client devices.

Advantages:

- Ease of deployment: NAS devices are relatively easy to deploy and configure, making them suitable for small to medium-sized organizations with limited IT resources.
- Shared storage: NAS devices allow multiple users and client devices to access shared storage resources over a network, facilitating collaboration and data sharing.
- Data redundancy: Many NAS systems support RAID (Redundant Array of Independent Disks) configurations for data redundancy, helping protect against data loss in the event of disk failures.

Disadvantages:

- Limited scalability: NAS devices may have limitations in terms of scalability compared to cloud storage solutions, making them less suitable for organizations with rapidly growing data storage needs.
- Network dependency: NAS devices rely on network connectivity, and performance may be impacted by network congestion or interruptions.
- Single point of failure: A single NAS device represents a single point of failure, and organizations need to implement backup and disaster recovery strategies to mitigate this risk.

3.2.2.2.4. External Hard Drives with Encryption

External hard drives equipped with encryption capabilities provide portable storage solutions for research data while ensuring data confidentiality. Encryption standards such as FIPS 140-2 (USA Federal Information Processing Standard) specify requirements for cryptographic modules used in hardware devices, including external hard drives. Adherence to such standards ensures that encryption mechanisms meet recognized security benchmarks. Encryption tools such as BitLocker

(for Windows OS), FileVault2 (for macOS), PGP, VeraCrypt, Axcrypt, and SafeHouse can be used if appropriate.

Advantages:

- Portability: External hard drives offer portable storage solutions, allowing researchers to transport data between different locations or share data with collaborators easily.
- Data encryption: External hard drives with encryption provide an additional layer of security for stored data, protecting against unauthorized access in case of loss or theft.
- Cost-effectiveness: External hard drives are generally cost-effective storage solutions, particularly for smaller-scale research projects with limited budgets.

Disadvantages:

- Physical vulnerability: External hard drives are susceptible to physical damage, loss, or theft, which can compromise the security and integrity of stored data.
- Limited capacity: The storage capacity of external hard drives is typically limited compared to other storage solutions, which may be insufficient for large-scale research projects or datasets.
- Manual backups: Researchers need to manually backup data to external hard drives, which can be time-consuming and prone to human error if not done regularly.

3.2.2.3. Establish clear data retention policies specifying retention periods, data categories, responsible parties, and procedures for data disposal.

Determine the purpose of data collection and retention for each research project. Retain data only for as long as necessary to achieve research objectives and document the rationale for retention periods. Comply with GDPR, national, and regional data protection laws, dictating data retention periods. Communicate policies to all stakeholders and ensure compliance through regular audits and monitoring.

Outline procedures for securely disposing of research data once it is no longer needed, including data deletion or archival processes considering the following steps: Data Assessment, Data Deletion, Archival Processes, Data Destruction, Documentation, Verification, Training and Awareness. Ordinary file and folder deletion using standard methods may not permanently erase the data. With the assistance of specialized tools, it is possible to recover them. There are software solutions available for permanent deletion, including Eraser, WipeFile, and FreeRaiser.

3.2.3. Access Control Procedures

- Define roles and permissions for accessing research data, including who has access to raw data, processed data, and analytical tools. For example, raw data, which is often the most sensitive and vulnerable to manipulation, should only be accessible to authorized personnel directly involved in data collection and processing. Processed data, which has undergone some level of transformation or analysis, may be accessible to a broader group of researchers or collaborators based on their roles and project requirements. Analytical tools, such as software applications or databases, should have restricted access based on the specific needs of users and their level of expertise.

- Implement authentication mechanisms, such as passwords or multi-factor authentication, to control access to research data.
- Specify procedures for granting and revoking access rights to research data, including documenting access requests and approvals. Also, organizations may choose to define access privileges or other attributes by account, type of account, or a combination of both. System account types include individual, shared, group, system, anonymous, guest, emergency, developer, manufacturer, vendor, and temporary. Other attributes required for authorizing access include restrictions on time-of-day, day-of-week, and point-of-origin. In defining other account attributes, organizations consider system-related requirements (e.g., system upgrades scheduled maintenance,) and mission or business requirements, (e.g., time zone differences, customer requirements, remote access to support travel requirements) (NIST SP 800-171).
- Monitor and control remote access sessions. Remote access is access to organizational systems by users (or processes acting on behalf of users) communicating through external networks (e.g., the Internet). Remote access methods include dial-up, broadband, and wireless. Organizations often employ encrypted virtual private networks (VPNs) to enhance confidentiality over remote connections. The use of encrypted VPNs does not make the access non-remote; however, the use of VPNs, when adequately provisioned with appropriate control (e.g., employing encryption techniques for confidentiality protection), may provide sufficient assurance to the organization that it can effectively treat such connections as internal networks. VPNs with encrypted tunnels can affect the capability to adequately monitor network communications traffic for malicious code (NIST SP 800-171). Automated monitoring and control of remote access sessions allows organizations to detect cyber-attacks and help to ensure ongoing compliance with remote access policies by auditing connection activities of remote users on a variety of system components (e.g., servers, workstations, notebook computers, smart phones, and tablets). (SP 800-46), (SP 800-77), and (SP 800-113) provide guidance on secure remote access and virtual private networks.
- Verify and control/limit connections to and use of external systems. External systems are systems or components of systems for which organizations typically have no direct supervision and authority over the application of security requirements and controls or the determination of the effectiveness of implemented controls on those systems. External systems include personally owned systems, components, or devices and privately-owned computing and communications devices resident in commercial or public facilities. This requirement also addresses the use of external systems for the processing, storage, or transmission of information, including accessing cloud services (e.g., infrastructure as a service, platform as a service, or software as a service) from organizational systems. Organizations establish terms and conditions for the use of external systems in accordance with organizational security policies and procedures. Terms and conditions address as a minimum, the types of applications that can be accessed on organizational systems from external systems. If terms and conditions with the owners of external systems cannot be established, organizations may impose restrictions on organizational personnel using those external systems. This requirement recognizes that there are circumstances where individuals using external systems (e.g., contractors, coalition partners) need to access organizational systems. In those situations, organizations need confidence that the external systems contain the necessary controls so as not to compromise, damage, or otherwise harm organizational systems.

Verification that the required controls have been effectively implemented can be achieved by third-party, independent assessments, attestations, or other means, depending on the assurance or confidence level required by organizations(NIST SP 800-171).

- Protect wireless access and control connection of mobile devices. Before allowing wireless connections, organizations establish usage restrictions and configuration requirements to authorize access to their systems, reducing the risk of unauthorized entry through wireless technologies. Authentication protocols used in wireless networks protect credentials and enable mutual authentication. Special attention is given to authenticating individuals and devices, especially with the proliferation of Internet of Things devices that may have wireless access to organizational systems. Additionally, organizations control the connection of mobile devices, considering their diverse technical characteristics and capabilities. Measures include device identification, authentication, configuration management, implementation of protective software, scanning for malicious code, updating virus protection, and ensuring software integrity, among others. Further guidance on securing mobile devices is provided in [SP 800-124].

Two illustrative examples of implemented case studies within the project highlight informative, non-normative approaches to data security:

Case[23] 1:

> Data generated during the project will be stored at the servers of the partner institutions located in physically secured environments. The computers, laptops, internet, or hard drives that will be used for processing and analysis of various data will be accessible through institutional passwords periodically modified according to national law for data security and protected by regularly updated antiviruses. None of the project data will be left inadvertently available. Data will be subject to regular backup to safeguard them from accidental losses. The project will be conducted in compliance with the EU General Data Protection Regulation (GDPR). When researchers from several partner institutions are working on the same task, and if the internal storage capacities and procedures of a partner prevent access to project staff from other partner institutions, the data will be temporarily transferred via a secure FTP protocol, as recommended by the IT team, so that the authorized researcher could finally transfer to storage. Additional options are to temporarily use some of the recommended cloud (web-based) storage (e.g. Amazon Web Services, Microsoft Azure or Google Suite for Education). It should be noted that the use of Dropbox, personal Google accounts (@google.com), personal OneDrive accounts, Apple iCloud is not recommended.

Case 2:

> If the project is classified as sensitive (if, for example, it includes electronic healthcare records, genomic data, and clinical images), certain additional data security measures or strict protocols are followed; e.g., the project will be conducted only after approval of the local ethical committee and adhering strictly to its regulations; fully compliant with European and international laws and policies governing research

[23] https://cordis.europa.eu/project/id/101046758/results

involving human patients: the Helsinki declaration in its latest version; the Charter of Fundamental Rights of the EU (2000/C 364/01); EU General Data Protection Regulation n. 2016/679 and national implementations, also for re-use of previously collected data; the European directive 95/46 EEC on the protection of individuals with regard to the processing of personal data; the principles laid down in the Oviedo Bioethics Convention. In addition, during the project, the following should be collected: a written informed consent form; written information to be provided to patients; clinical sites ethics committees' approval.

In terms of data management for a project of this nature, the recommended protocol entails:

Within the institutional server, there should be a folder specifically allocated for the project. A laptop from the project lab, protected with a password, will be connected to this server. The PI and two other authorized individuals should have access to it, also requiring a password for entry. Data from previously processed patients will be retrieved from an external hard drive, undergo pseudonymization on the laptop, and then be transmitted to the server. To grant access to this data for other project partners, whenever someone attempts to access the database, the PI will receive a notification and determine whether to grant or deny access.

3.2.4. Data Sharing Procedures

- Establish protocols for sharing research data with collaborators, funders, or other stakeholders, ensuring compliance with data sharing agreements and confidentiality requirements. One common protocol is to use secure data sharing platforms or repositories that adhere to industry standards such as ISO/IEC 27001 or NIST SP 800-171. For example, platforms like Globus, Figshare, or Zenodo provide secure environments for sharing research data while offering encryption, access controls, and audit trails to protect data integrity and confidentiality.
- Define the procedures for anonymizing or aggregating data when sharing sensitive information to protect individual privacy. (Section 4.4).
- When sharing research data, especially with collaborators, funders, or other institutions, it's crucial to establish clear agreements to govern the transfer of data and any associated materials. Two common types of agreements used in this context are Data Transfer Agreements (DTAs) and Material Transfer Agreements (MTAs). A Data Transfer Agreement (DTA) is a legal contract outlining terms for sharing research data between parties, specifying rights, responsibilities, and data usage, while a Material Transfer Agreement (MTA) governs the transfer of tangible research materials, establishing terms for their use, ownership, and distribution between organizations or institutions.
- The principle of controlling the flow of information involves regulating information movement within and between systems, regardless of who accesses it. This regulation contains various measures such as preventing export-controlled data from being transmitted over the internet in plain text, blocking external traffic posing as internal, restricting internet requests not routed through internal proxies, and managing information transfers between organizations based on content and structure. Organizations typically implement policies and enforcement mechanisms to manage information flow, relying on the characteristics of the data or its path.

Enforcement occurs in boundary protection devices (e.g., gateways, routers, guards, encrypted tunnels, firewalls) that employ rule sets or establish configuration settings that restrict system services, provide a packet-filtering capability based on header information, or message-filtering capability based on message content (e.g., implementing key word searches or using document characteristics). To mitigate risks when transferring information between systems with different security policies, organizations employ guidance at policy enforcement points and may mandate architectural solutions to enforce specific security policies, including prohibiting transfers, employing one-way information flows, and implementing trustworthy regrading mechanisms to adjust security attributes. Additionally, separating duties among individuals helps mitigate the risk of malicious activity by dividing mission and support functions among different roles and ensuring that personnel administering access control functions are distinct from those managing audit functions. This approach spans systems and application domains to address potential security vulnerabilities comprehensively.

3.2.5. Data Security Incident Response Procedures

Incident - An occurrence that actually or potentially:

- Jeopardizes the confidentiality, integrity, or availability of an information system
- Jeopardizes the information the system processes, stores, or transmits
- Constitutes a violation or imminent threat of violation of security policies, security procedures, or acceptable use policies.

Breach - An incident that results in a confirmed disclosure of data to an unauthorized party.

The academic sector is facing a serious cybersecurity challenge in 2024, with more than 32% of academic entities reporting cyber attacks in 2023 alone(Windows Management Experts INC.). Higher education had the highest rate of ransomware attacks among all industries surveyed in a 2016 report published by BitSight (a cyber risk management company), and the second highest rate in BitSight's 2017 report. In 2015, a U.S. university stated it faced 20 million cyber attacks per day[24]. The National Security Agency (NSA) warned in its *Cybersecurity Year in Review* report that academic research institutions are prime targets for cyber espionage. Faculties and research organizations manage vast amounts of sensitive data, ranging from student personal data to valuable intellectual property. If this data were to be stolen or compromised, it could lead to significant repercussions extending well beyond the institution itself. Beyond potential financial losses, cyberattacks present a serious risk to a university's reputation, clandestine intellectual property acquisition and the safety of its students.

Hence, it's crucial that that higher academia and research institutions (in particular, University of Belgrade) prioritize reducing the frequency of incidents by effectively securing networks, systems, and applications while maintaining up-to-date documentation to support incident handling.

3.2.5.1. Pre-incident actions

- Establish policies for internal and external reporting

[24] https://archive.nytimes.com/bits.blogs.nytimes.com/2015/05/15/penn-states-college-of-engineering-hit-by-cyberattack/?_r=0

Establish a regular schedule to refresh training, software/firewall/antivirus updates, and policy reviews. In safeguarding digital assets, organizations deploy various protective measures across different levels of their IT infrastructure. Host-based protection software includes antivirus software, Host Intrusion Detection Systems (HIDS), Host Intrusion Prevention Systems (HIPS), and Whitelists, all geared toward fortifying individual devices against cyber threats. Complementing this, network-based protection software includes Firewalls, Network Intrusion Detection Systems (NIDS), and Network Intrusion Prevention Systems (NIPS), which secure the broader network environment, thereby forming a comprehensive defense mechanism against potential cyberattacks.

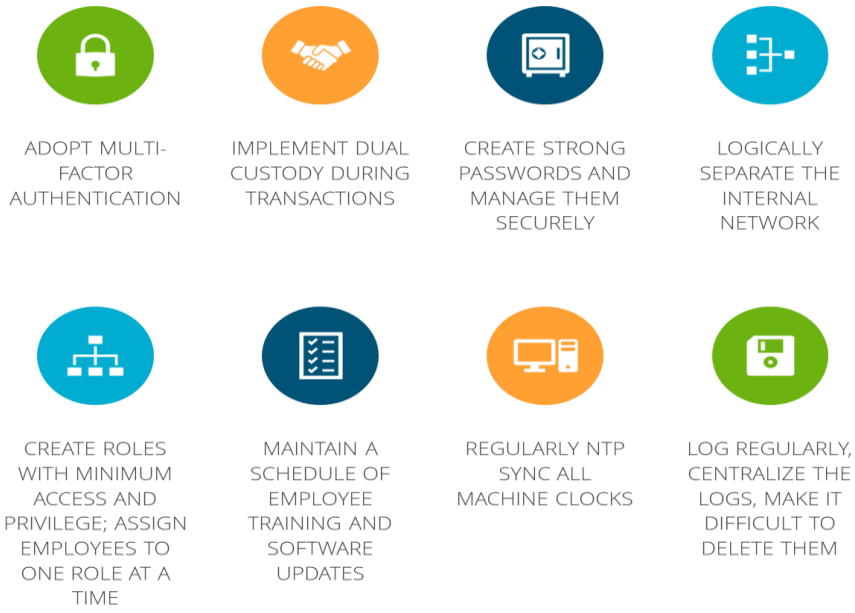

Figure 3: Best-Practice Policies for a Layered Defense

3.2.5.2. Incident and post-incident actions

Develop a protocol for responding to data breaches or security incidents, including reporting procedures and escalation paths. The procedure ought to include several of the subsequent steps:

- Detection and Assessment: Identify and verify the security incident or data breach. Assess the scope, severity, and potential impact of the incident.
- Containment and Mitigation: Take immediate action to contain the breach or incident to prevent further unauthorized access or damage. This may involve isolating affected systems or networks.
- Notification: Notify relevant stakeholders, including internal teams, management, affected individuals, and regulatory authorities, as required by applicable laws and regulations. Understand legal requirements for reporting. For example, if subject to the EU General Data Protection Regulation (GDPR), the law requires a data breach notification to the responsible national supervisory authorities without undue delay, that is not later than 72 hours after awareness of the incident. Provide clear and

timely communication about the incident, its impact, and any measures being taken to address it.

- Investigation and Analysis: Conduct a thorough investigation to understand the root cause of the incident, identify vulnerabilities or weaknesses in security controls, and determine the extent of data exposure or compromise.
- Remediation and Recovery: Implement remediation measures to address the vulnerabilities or weaknesses identified during the investigation. Restore affected systems or data to a secure state and ensure continuity of operations.
- Documentation and Reporting: Document all aspects of the incident response process, including actions taken, findings from the investigation, remediation efforts, and lessons learned. Report the incident to relevant authorities, such as data protection authorities or law enforcement, as required by law.
- Review and Lessons Learned: Conduct a post-incident review to evaluate the effectiveness of the response process, identify areas for improvement, and update incident response plans and security controls accordingly.
- Outline steps for investigating security incidents, mitigating risks, and implementing corrective actions to prevent recurrence.

3.2.6. User Training and Awareness

Research organizations cannot protect the confidentiality, integrity, and availability of information in today's highly networked systems environment without ensuring that all people involved in using and managing IT: understand their roles and responsibilities related to the organizational mission; understand the organization's IT security policy, procedures, and practices; and have at least adequate knowledge of the various management, operational, and technical aspects. Incorporating robust security awareness and training programs customized to the unique requirements of the organization and the roles of individuals is essential for effectively mitigating security risks and fostering a culture of security awareness within the organization. These include ensuring that managers, systems administrators, and users are aware of security risks and related policies, standards, and procedures and that personnel are trained to carry out their assigned information security-related duties and responsibilities. Organizations tailor security awareness training content and techniques to specific requirements and access levels, covering information security fundamentals, user actions, and operations security. Techniques include formal training, email advisories, logon screen messages, security awareness posters, and events. Comprehensive role-based training covers management, operational, and technical roles, including physical, personnel, and technical controls. Training also addresses insider threat awareness, emphasizing recognition and reporting of potential indicators through appropriate channels.

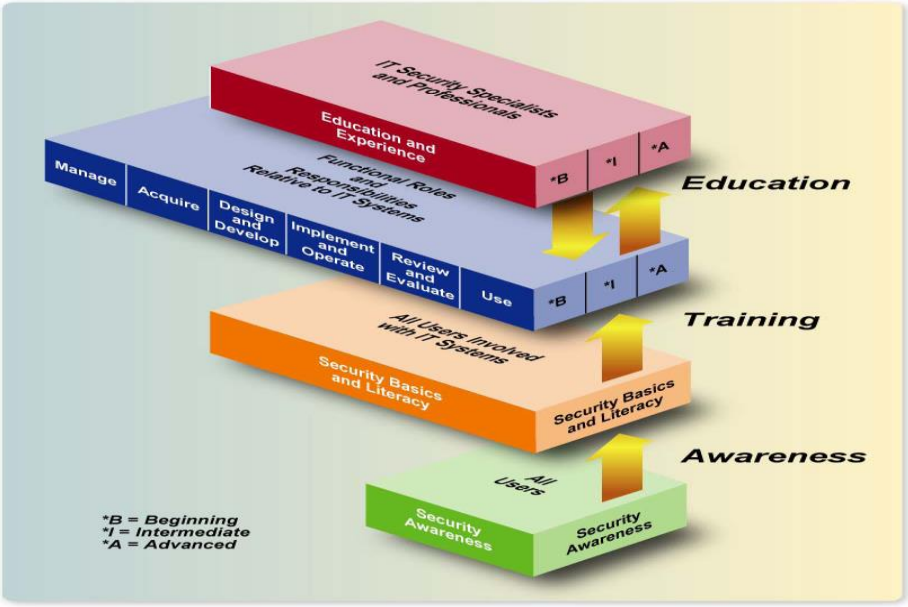

Figure 4: The IT Security Learning Continuum (SP 800-50)

Additional guidance is provided by standards such as SP 800-50, SP 800-181, and SP 800-161, offering insights into security awareness and training programs, role-based training, and supply chain risk management. Moreover, Appendix I incorporates sample metric questions tailored to aid in assessing security awareness and training efforts.

3.3. List of Available Tools

- **Adobe Acrobat** — program for creating, viewing, and managing PDF documents
- **Adobe Bridge** — program for managing digital assets
- **Bulk Rename Utility** — tool for automatically renaming files through software solutions
- **GitHub** — online tool for version control of data
- **Microsoft File Checksum Integrity Verifier**— tool for checking the integrity of digital files
- File encryption programs for files, folders, portable media, and hard drives:
  - **Axcrypt**
  - **BitLocker** for Windows OS
  - **FileVault2** for macOS
  - **PGP**
  - **SafeHouse**
  - **VeraCrypt**
- Programs for permanent file deletion:
  - **Eraser**
  - **FreeRaiser**
  - **WipeFile**
- **TagSpaces** — software solution for file management

4. AFTER THE RESEARCH (LIFECYCLE: PERMANENT STORAGE, SHARING, UTILIZATION)

This chapter provides an overview of the last three phases in the lifecycle of research data, focusing on permanent storage, sharing, and utilization of the data collected and processed during the project. Furthermore, the topics covered in this chapter include types of licenses, the process of selecting an appropriate license, and the management and publication of sensitive data, which often involve anonymization and/or pseudonymization procedures.

4.1. Repositories

A 'repository' for scientific publications is an online archive where researchers deposit research publications, research data and other scientific R&D outputs, and can preserve, manage and provide access to them. They collect and disseminate digital research outputs from individual research organizations' institutional repositories (e.g., the repository of the Belgrade University) or specific research communities' thematic/disciplinary repositories (e.g., Europe PMC for life sciences including biomedicine and health or arXiv for physics, mathematics, computer science, quantitative biology, quantitative finance and statistics, etc.). They can also be centralized repositories, such as for example Zenodo, free of charge and developed by CERN.

The beneficiaries can identify and choose the repository that suits their needs in the Directory of Open Access Repositories (Open DOAR) (Please see: Appendix H). The Open Access Infrastructure for Research in Europe (OpenAIRE) can help them to find such repositories. The beneficiaries are recommended to use a repository that is compliant with the requirements of OpenAIRE and to use the OpenAIRE database as a point of departure. The OpenAIRE also offers support services for researchers by providing information on open access requirements, a general Helpdesk and country-specific helpdesks (National Open Access Desks) that answer questions on open access and GA project. Beneficiaries should NOT use a repository with rules that could conflict with open access requirements.

The underlying data needed to validate the results presented in scientific publications is seen as a crucial part of the publication and therefore an important element of scientific R&D best practice. For ease of tracking, beneficiaries should also include the digital object identifier (DOI) for the GA project (http://dx.doi.org/10.13039/501100007601) in the funding acknowledgement field in their metadata. The metadata compliance of the repository can be checked using OpenAIRE.

It is important to check whether the funders have already predetermined a repository for long-term data storage or have set criteria that the repository must meet. The recommendation is to store research data in the institutional repository, which supports FAIR principles and enables publication in accordance with the requirements set by the European Commission. Searching thematic repositories is possible through re3data, a registry that gathers more than 2000 repositories. If the data are stored in a general repository, the recommendation is for it to be Zenodo.

After depositing publications, beneficiaries must ensure open access to those publications via the chosen repository, choosing one of two main ways to meet this requirement:

1. Open access publishing or 'gold' OA: Researchers can publish in open access journals' repository or in hybrid journals in which both options, selling subscriptions or making individual articles openly accessible, are accepted. In open access mode, an article is immediately published and the payment of publication costs is shifted away from subscribing readers. The most common business model is based on one-off payments by authors. These costs, often referred

to as Article Processing Charges (APCs), are usually borne by the researcher's university or research institute or the agency funding the research. Monographs can also be published either on a purely open access basis or using a hybrid business model. 'Article processing charges' are eligible for reimbursement during the duration of the project. The costs of 'gold' open access publications, incurred once a project is completed, cannot be refunded from that project's budget.

2. Self-archiving or 'green' OA: The beneficiaries or representatives can deposit the final peer-reviewed manuscript into a repository of their choice, before or at the same time, or after publication (within at most 6 or 12 months). Some publishers request that open access be granted only after an embargo period has elapsed. To provide support concerning compliance with the GA project embargo periods, the EU Commission offers a model amendment to publishing open access agreement to be filled out by the beneficiary and the publisher. This model is not mandatory but reflects the obligations for the beneficiary under the GA. It can be supplemented by further provisions agreed between the parties, provided they are compatible with the GA. The Commission/Agency takes no responsibility for the use of this model which is often signed between authors and publishers.

However, not all data can be open, meaning that research data must be "as open as possible and as closed as necessary". Partially or entirely open access to deposited research data has become the default setting for research data generated in the GA projects. Beneficiaries can therefore opt out at any stage of projects, either before or after signing the GA, and so free themselves retroactively from the obligations associated with the conditions, if:

- opening access is incompatible with the obligation to protect results that can reasonably be expected to be commercially or industrially exploited,
- opening access is incompatible with the need for confidentiality and security requirements,
- opening access may reasonably provide data with dual-use applications,
- opening access is incompatible with rules on protecting personal data,
- opening access would mean that the project's main aim might not be achieved, or the project will not generate/collect any research data, or there are other legitimate interest reasons (e.g., that beneficiary can enter in a free-text box at the proposal stage for the GA and explain it).

Alternatively, any project under the GA can also choose to keep selected datasets or even all research data closed for any of the above reasons via their Data Management Plan.

4.2. Types of Licenses

Creative Commons[25] (CC) is the most popular licensing system widely accepted since 2001 and continuously improved through versioning. There are four main conditions defining the use of content:

1. Attribution (BY) — modification, adaptation, and sharing of the work are permitted with attribution to the original author.

[25] https://creativecommons.org/share-your-work/cclicenses/

2. NonCommercial (NC) — modification, adaptation, and sharing are permitted exclusively for non-commercial purposes.
3. NoDerivatives (ND) — the work can be used only in its original form, and no modifications are allowed.
4. ShareAlike (SA) — modification, adaptation, and sharing are permitted under the same conditions (license) as the original work.

By combining these conditions, Creative Commons has developed six licenses:

1. CC BY (Attribution)
2. CC BY-SA (Attribution-ShareAlike)
3. CC BY-ND (Attribution-NoDerivatives)
4. CC BY-NC (Attribution-NonCommercial)
5. CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
6. CC BY-NC-ND (Attribution-NonCommercial-NoDerivatives).

4.2.1. Which licenses should be used for research data?

4.2.1.1. Public Domain (CC0)

Generally, the less restrictive the license, the greater the possibility of reuse and correct usage. CC0, or the dedication of works to the public domain, allows others to reuse the data without restrictions. While attribution is not legally required when using data published under the CC0 license, citing is a norm imposed by scientific integrity.

4.2.1.2. Attribution (CC BY 4.0)

CC BY 4.0 allows for modification, adaptation, and sharing with attribution. It is necessary to credit the author, provide a link to the original work (dataset), and indicate any changes made. When using CC BY licenses in the context of data, caution should be exercised due to the potential accumulation of attributions. Using CC0 with a request for attribution is the best option. Alongside published data, there may be a copyright notice stating "Please attribute my data" or a similar phrase. The use of CC0 and CC BY 4.0 licenses is recommended by the European Commission for Horizon 2020 projects and RDA.

Once a CC license is assigned, it cannot be changed, so caution should be exercised when selecting. If it is not possible to clearly separate the use of different datasets with different licenses in a new dataset; the most restrictive license must be applied to it. When publishing a dataset, it is necessary to clearly indicate the license under which the data are published. In repositories and archives, the license must be selected during the data storage process. Additionally, it is advisable to specify the license in the readme file.

4.3. Informed Consent

Before participating in the research, participants are provided with informed consent, which they sign to give the researcher permission to process and publish the data collected for research purposes. In the consent form, the researcher must inform participants about the purpose of the research, methods of data collection and analysis, forms of data dissemination and publication, and data management after the project ends. To enable the reuse of research data, the researcher should clearly state the possibility of future use. It is important to specify how data protection will be ensured, for example, through anonymization procedures. Additionally, the right of participants to withdraw from the study should be highlighted. It is not recommended to use phrases such as "completely anonymous" or "strictly confidential" because achieving such levels of anonymity or confidentiality is practically impossible. Also, promises of data destruction or limiting data access only to the research team should be avoided.

The consent form should be written in clear and understandable language. According to the General Data Protection Regulation [26](GDPR), for consent to be valid, the participant must voluntarily give specific, informed, and unambiguous consent to data processing, either orally or in writing. An example of a consent form for the collection, processing, and publication of personal and sensitive data for scientific research can be seen in Elsevier's guidance. Informed consent protects all parties involved in the research (participants and researchers) and shapes the research process. In addition to informed consent, researchers should provide participants with a GDPR statement if they will be collecting personal data during the research.

4.4. Anonymization and Pseudonymization

Sensitive and personal data need to be anonymized to protect the identity of respondents. There are two types of data anonymization — anonymization and pseudonymization — and both processes can be applied to qualitative and quantitative data. Quantitative data include numerical and alphanumeric values, variables, and attributes, while qualitative data encompass textual data related to interview transcripts, audio and video recordings, and visual materials. ENISA (The European Union Agency for Cybersecurity) provides definitions for anonymization and pseudonymization. Anonymization is defined as an irreversible process of modifying personal data after which the individual can no longer be directly or indirectly identified. The anonymization process involves removing direct and/or indirect identifiers from the dataset. Data may be altered or reorganized to prevent the disclosure of respondents' identities. The anonymization process cannot be reversed; it permanently anonymizes personal and sensitive data of respondents. Pseudonymization, as defined by ENISA, is a data protection technique where personally identifiable information (PII) within a dataset is replaced by artificial identifiers, or pseudonyms. These pseudonyms allow data to be processed and analyzed without revealing the individual's identity directly. However, if necessary, additional information can be used to re-identify individuals.

Notions[27] to keep in mind: pseudonymization vs. anonymization.

Pseudonymization still makes the data subject identifiable, through the combination of the pseudonym (e.g. key-code, code number) with additional identifiers. The time and the effort required to identify the individual as well as the available technologies are decisive for determining

[26] Condition for consent https://gdpr-info.eu/art-7-gdpr/
[27] https://www.eui.eu/documents/servicesadmin/deanofstudies/researchethics/guide-data-protection-research.pdf

whether is possible to identify the data subject from pseudonymized data. Anonymization excludes any possibility to identify the data subject.

4.5. Long-Term Storage of Final Research Results

Data constitutes the essence of scientific research, underscoring the paramount importance of data security. The repercussions of data loss can be severe, with recovery efforts often proving to be slow, costly, or even futile. Thus, decisions regarding data security and storage are pivotal facets of effective data management. Within the institution, the IT department should be tasked with managing reliable backups to mitigate risks associated with hardware failures, software errors, cyber threats, power outages, and human errors. Additionally, scientists in research organizations can ensure cybersecurity by engaging in education, collaborating with IT teams, adopting secure technologies, complying with regulations, conducting risk assessments, and maintaining awareness of emerging threats.

The institution's network server boasts robust infrastructure for securely storing data, making it a primary choice for implementing backup strategies. Network servers offer numerous advantages, including regular maintenance, minimal risk of failure or unauthorized access, and the expertise of network administrators readily available for support (Section: 3.2.2.2.). Furthermore, cloud services are favored for their simplicity and accessibility. Reliance solely on computer hard disks is discouraged, especially for storing master copies of research. Portable storage media like CDs, DVDs, and flash drives are convenient for mobility but are not recommended for long-term storage due to susceptibility to damage, loss, or theft, as well as the risk of digital obsolescence. To ensure a successful data management strategy, adherence to backup best practices is advised. This includes designating a responsible individual for regular backups, following the 3-2-1 rule (making three copies on two different devices, with one stored offsite), adhering to a backup schedule, selectively backing up relevant data, and tracking versions to maintain data integrity.

Document all procedures related to data storage and preservation in the README file, such as where the data are stored, how many copies, in which formats, and so on. Certain types of data and sensitive research may necessitate specific storage restrictions, with the choice of storage solution greatly influencing collaboration and version control. While project funders may dictate conditions for data storage, adhering to principles of necessity and redundancy – storing only essential data and maintaining multiple copies in diverse locations – is recommended in the absence of explicit guidelines.

If the project possesses an innovative nature and the potential for commercially exploitable results requiring intellectual property protection (such as patents, NDAs, and trade secrets), certain datasets will need to be restricted or accessible only to authorized users (with contact information provided). However, the metadata will be licensed under CC0.

Each beneficiary must examine the possibility of protecting its results and must adequately protect them, for an appropriate period and with appropriate territorial, security and legal coverage, if:

> a. the results can reasonably be expected to be commercially or industrially exploited, and
> b. protecting them is reasonable and justified, in given circumstances, even if this requires further research and development or private investment.

The beneficiaries are in principle free to choose any available standard form of protection, such as: (1) Patent; (2) Trademark; (3) Industrial design; (4) Copyright; (5) Trade-secret, and (5) Confidentiality (e.g., information security and privacy protection).

If the intellectual property generated stems from research involving a single partner institution, that institution, being the data owner, will determine when and if the data can be opened, considering its legitimate interests and internal processes. In cases where multiple partners contribute to the intellectual property, the Consortium Agreement will govern, with any undefined matters addressed in a new Joint IP Ownership and Management Agreement. This agreement will outline which datasets will be restricted and for how long.

Further details on data management should be provided in the final version of the Data Management Plan, covering:

- Criteria for selecting datasets for long-term preservation
- Obligations regarding data retention or destruction for contractual, legal, or regulatory compliance
- Anticipated research uses and potential users of the data
- Required tools for accessing and utilizing the data
- Sustainability measures for the software necessary to access the data.

4.6. List of Available Tools

- **Amnesia** — online tool for anonymizing quantitative data stored in .csv and .txt formats
- **COPTR (Community Owned Digital Preservation Tool Registry)**— online tool assisting researchers in finding data preservation tools, i.e., a registry of data storage tools
- **Creative Commons (CC)** — licensing system
- **RDM Advice & Tips** — advice on privacy, anonymization, and pseudonymization of research data
- **Re3data** — enables searching thematic repositories, gathering more than 2000 repositories
- **UK Data** — advice for anonymizing quantitative and qualitative data
- **Zenodo** — general data repository.

APPENDIX A. PROJECT DATASETS

No	Dataset Title	Responsible Person/Partner	Short Description	Format	Open Access	Closed/Restricted	Repository/PD
					License (CC BY or CC0)	Justification	
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							

APPENDIX B. GUIDELINE QUESTIONS FOR DEVELOPING DATA SUMMARY SECTION OF DMP

Instructions for answering the questions are highlighted in yellow.

1. What types of data will the project generate or reuse (e.g., experimental data, observational data, images, text, biomaterial…)?

   A:

2. What data (for example the kinds, formats, and volumes) will be collected or produced?
   a. Give details on the kind of data: for example, numeric (databases, spreadsheets), textual (documents), image, audio, video, and/or mixed media.
   b. Give details on the data format: the way in which the data is encoded for storage, often reflected by the filename extension (for example pdf, xls, doc, txt, or rdf).
   c. All possible formats could be: PDF/A (.pdf) and ODT (.odt) for textual files; NetCDF and TextFabric for programs used while processing data; ODS (.ods) and CSV (.csv) for charts; SQL (,sql), SIARD (.siard), JSON (.json) and CSV (.CSV) for data bases; SPSS (.dat/sps), STATA (.dat/.DO) and R for statistic data; T1FF (.tif,.tiff,), PNG(.png), JPEG, JPEG 2000 (.jp2) and DICOM (.dcm) for images; SVG (.SVG) for vector images and other electronic images; BWF (.bwf), MXF (.mf), Matroska (.mka), FLAC (.flac) and OPUS for audio materials; MHF (.mhf) and Matroska (.mкv) for video materials; WaveFront Object (.obj), Polygon file format (.p1y), X3D (.x3d) and COLLADA (.dae) for 3D images; RDF/XML (.rdf), Trig (.trig), Turtle (.ttl), NTriples (.nt) and JSON-LD for graphs with tags; REFI-QDA (Qualitative Data Analysis) for qualitative computer analyses.
   d. Justify the use of certain formats: For example, decisions may be based on staff expertise within the host organization, a preference for open formats, standards accepted by data repositories, widespread usage within the research community, or on the software or equipment that will be used.
   e. Give preference to open and standard formats as they facilitate sharing and long-term re-use of data (several repositories provide lists of such 'preferred formats').
   f. Give details on the volumes for each type, they can be expressed in storage space required (bytes), and/or in numbers of objects, files, rows, samples and columns (e.g., experimental data……. MB /GB, observational data……. MB /GB, images……. MB, /GB, text…… MB /GB, biomaterial… (Samples…), genetic data samples …).).

   A:

3. What is the purpose of generated or reused data?

   A:

4. What is the origin/provenance of the data, either generated or reused data, and its relation to the objectives of the project?

   A:

5. Explain how data provenance will be documented?

   A:

6. Will you reuse any existing data through the project and for what purpose?

   A:

7. What are the reasons for reuse of any existing data sources considered but discarded (if any)?

   A:

8. What is the potential of the research data to be used outside of the project and to whom might your data be useful ('data utility'), outside your project?
   A:
9. How will new data be collected or produced and/or how will existing data be reused?
   A:
10. Explain which methodologies or software will be used if new data are collected or produced?
    A:
11. State any constraints on reuse of existing data if there are any?

APPENDIX C. GUIDELINE QUESTIONS FOR DEVELOPING THE DATA FAIR VERIFICATION SECTION OF DMP

Instructions for answering the questions are highlighted in yellow.
This section of the DMP should present measures to ensure the data's Findability, Accessibility and Interoperability and Reusability.

C.1. Making data findable, including provisions for metadata

The first step in reusing data is to find them. Metadata and data should be easy to find for both humans and computers. Machine readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIR process verification (FAIRification). The data should be accurately described with rich metadata. The metadata should document how the data were generated, under what license and how they can be re-used, and provide the context for proper interpretation by other researchers. Including any identifiers, keywords, metadata standards and other practices that will optimize the potential of finding and re-using the data. In its most basic sense, metadata is information about data, and describes basic characteristics of the data, such as: who created the data, what the data file contains, when the data were generated, where the data were generated, why the data were generated, and how the data were generated. Metadata may also be used to document technical specifications of the files, relationships between multiple files, and rights and access information for the files. By providing metadata about the research files you wish to preserve, you will be making it easier for yourself and others to identify and reuse these data correctly at a later date. Metadata is "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource" (such as a data set). "A metadata record is a file of information which captures the basic characteristics of a data or information resource."

Note that there are many different metadata standards that you could use for assigning metadata to your files. Use the standard recommended by your funding opportunity specifications or your specific NSF directorate. If no guidelines are given, there are no standards for your community, or you find the standards inadequate, explain this situation and propose solutions or remedies. Even if standards are not available, you should at the very least include with your data a text document, such as a readme.txt" file, that includes the relevant information described above or references a published article that describes the data fully. Controlled vocabularies and ontologies exist for many fields of research. The data should be accurately described with rich metadata. The metadata should document how the data were generated, under what license and how they can be reused, and provide the context for proper interpretation by other researchers.

Data findable requires are:

F1: Metadata are assigned globally unique and persistent identifiers
F2: Data are described with rich metadata
F3: Metadata clearly and explicitly include the identifier of the data they describe
F4: Metadata are registered or indexed in a searchable resource

To achieve these the researchers must provide the following answers:

- Will you enable finding of data?
  A:
- Will data be identified by a persistent identifier?

A:

- Will rich metadata be provided to allow discovery?
  A:
- What metadata will be created?
  A:
- If there is not a standard format, how will you format your data so that others in your field will be able to make use of it?
  A:
- Are there any standard formats in your field for managing or disseminating the data sets you have identified (e.g., XML, ASCII, CSV, MySQL, net CDF)?
  A:
- If your format is proprietary rather than open, is this essential?
  A:
- Who on your team will have the responsibility for ensuring that data standards are properly applied and data are properly formatted?
  A:
- Are you aware of any metadata standards specific to your field that could be used for your data sets? (e.g., *Dublin Core [DC], Resource Description Format [RDF], Federal Geographic Data Committee [FGDC], Directory Interchange Format [DIF], Ecological Metadata Language [EML], Minimum Information About a Proteomics Experiment [MIAPE], and the Data Documentation Initiative [DDI]*)?
  A:
- Do you use metadata standards that are broadly accepted (e.g., *by the scientific community*)?
  A:
- Do you ensure that metadata are machine-retrievable?
  A:
- In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.
  A:
- Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential reuse?
  A:
- Will metadata be offered in such a way that it can be harvested and indexed?
  A:
- Do you enable referencing to related relevant information, such as other data and publications?
  A:
- Do you provide information that is publicly available and maintained, even for non-published, protected, retracted, or deleted data?
  A:
- How will metadata be generated and captured for each of your data sets?
  A:
- If there is not a metadata standard, what metadata will you need to generate so that
- others in your field will be able to find, understand, and make use of your data?

A:

- Who on your research team will be responsible for ensuring metadata standards are followed?
  A:
- Will metadata be made openly available and licensed under a public domain dedication **CC0 license**, as per the Grant Agreement?
  A:
- Will metadata contain information to enable the user to access the data?
  A:
- Will metadata be guaranteed to remain available after data is no longer available?
  A:
- Will documentation or reference about any software be needed to access or read the data be included?
  A:

C.2. Making data openly accessible and repository

First, details on the repository in which the data will be deposited should be given. Second, the access to the data itself, including open access, access protocols and restrictions aspects. Third, issues relating to metadata accessibility and availability should be described. In the case of certain data or metadata that will not be shared – proper justification should be provided. Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorization.

A1. Metadata are retrievable by their identifier using a standardized communications protocol.
A1.1 The protocol is open, free, and universally implementable.
A1.2 The protocol allows for an authentication and authorization procedure, where necessary.
A2. Metadata are accessible, even when the data are no longer available.

To make data accessibility the beneficiaries have to answer the following questions:

- Are data accessibility provided for the data management life cycle (*Plan, Collect, Assure, Describe, Preserve, Integrate and Analyze*)?
  A:
- Will all data be made openly available?
  A:
- If an embargo is applied to give time to publish or seek protection of the intellectual property (*e.g., patents)*, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible?
  A:
- Will the data be accessible through a free and standardized access protocol?
  A:
- If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?
  A:
- How will the identity of the person accessing the data be ascertained?
  A:

- Is there a need for a data access committee (e.g., *to evaluate/approve access requests to personal/sensitive data*)?
  A:
- The following table provides guidance for the selection of trustworthy repositories by criteria structured according to four main topics (**Please se**e Appendix H)
- Will the data be deposited in a trusted repository?
  A:.
- Have you explored appropriate arrangements with the identified repository where your data will be deposited?
  A:
- Does the repository ensure that the data is assigned an identifier?
  A:
- Will the repository resolve the identifier to a digital object?
  A:

C.3. Making data interoperable and reusable

The vocabularies, standards, formats or methodologies that will be used to enable data exchange, re-use and interoperability. This sub-section should provide information on the expected documentation (e.g., explaining methodology, codebooks, variables). The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

I1. Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. Metadata use vocabularies that follow FAIR principles.
I3. Metadata include qualified references to other metadata. The ultimate goal of **FAIR** is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined indifferent settings.
R1. Metadata are richly described with a plurality of accurate and relevant attributes.
R1.1. Metadata are released with a clear and accessible data usage license.
R1.2. (Meta)data are associated with detailed provenance.
R1.3. Metadata meet domain-relevant community standards.

To make data interoperable and reusable the beneficiaries have to answer the following questions:

- What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines?
  A:
- In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?
  A:
- Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?
  A:

- Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?
  A:
- How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?
  A:
- Will your data be made freely available in the public domain to permit the widest re-use possible?
  A: Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?
  A:
- Will the data produced in the project be useable by third parties, in particular after the end of the project?
  A:.
- Will the provenance of the data be thoroughly documented using the appropriate standards?
  A:
- Describe all relevant data quality assurance processes.
  A:

Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

APPENDIX D. GUIDELINE QUESTIONS FOR DEVELOPING THE OTHER RESEARCH OUTPUTS SECTION OF THE DMP

Instructions for answering the questions are highlighted in yellow.

The management of other research outputs that are generated/re-used in the project (e.g., software, models, new materials) should be discussed and, when relevant, their compliance to the FAIR principles should be detailed. **Please answer the following questions:**

- In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (*e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.*).
  A:
- Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.
  A:

APPENDIX E. GUIDELINE QUESTIONS FOR DEVELOPING THE ALLOCATION OF RESOURCES SECTION OF THE DMP

Instructions for answering the questions are highlighted in yellow.

This section should include a discussion on the resources such as costs associated with compliance to the FAIR principles or who will be responsible for data management. Please answer the following questions:

- What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.)?
  A:
- How will these be covered?
  A:
- Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions). Who will be responsible for data management in your project? How will long term preservation be ensured?
  A:
- Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?
  A:

APPENDIX F. GUIDELINE QUESTIONS FOR DEVELOPING THE DATA SECURITY AND PRIVACY SECTION OF DMP

Instructions for answering the questions are highlighted in yellow.
Please answer the following questions:

- What provisions are or will be in place for data security, including data recovery as well as secure storage/archiving and transfer of sensitive data?
  A:
- Will the data be safely stored in trusted repositories for long term preservation and curation?
  A:
- If personal data are processed, how will compliance with legislation on personal data and on security be ensured?
  A:
- Ensure that when dealing with personal data, *Data protection laws* (for ***example GDPR/National DPA***) are complied with:
  A:
- *Gain informed consent for preservation and/or sharing of personal data.*
- *Consider anonymization of personal data for preservation and/or sharing (truly anonymous data are no longer considered personal data).*
- *Consider pseudonymization of personal data (the main difference with anonymization is that pseudonymization is reversible).*
- *Consider encryption which is seen as a special case of pseudonymization (the encryption key must be stored separately from the data, for instance by a trusted third party).*
- *Explain whether there is a managed access procedure in place for authorized users of personal data.*
- How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?
  A:
- Explain who will be the owner of the data, meaning who will have the rights to control access?:
- *Explain what access conditions will apply to the data? Will the data be openly accessible, or will there be access restrictions? In the latter case, which?*
- *Consider the use of data access and re-use licenses.*
- *Make sure to cover these matters of rights to control access to data for multi-partner projects and multiple data owners, in the consortium agreement.*
- *Indicate whether intellectual property rights (for example Database Directive, sui generis rights) are affected. If so, explain which and how will they be dealt with.*
- *Indicate whether there are any restrictions on the re-use of third-party data.*

APPENDIX G. GUIDELINE QUESTIONS FOR DEVELOPING THE ETHICS ISSUE SECTION OF THE DMP

Instructions for answering the questions are highlighted in yellow.

Any ethical or legal issues that can have an impact on data sharing should be presented. Additionally, when the research uses personal data, aspects such as informed consent or long-term preservation should be referred to. To achieve these the **researchers must provide the following answers**:

- What ethical issues and codes of conduct are there, and how will they be taken into account?
  A:
- Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?
  A:
- How will possible ethical issues be taken into account, and codes of conduct followed?
  A:
- If personal data are processed, how will compliance with legislation on personal data and on data security be ensured?
  A:
- Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?
  A:
- How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?
  A:
- Consider whether ethical issues can affect how data are stored and transferred, who can see or use them, and how long they are kept. Demonstrate awareness of these aspects and respective planning?
  A:
- Follow the national and international codes of conducts and institutional ethical guidelines, and check if ethical review (for example by an ethics committee) is required for data collection in the research project?
  A:

APPENDIX H. RECOMMENDED AND SELECTED REPOSITORIES AND DATABASES

- **Zenodo** - a general-purpose open research data repository for the preservation and making available of research, educational and informational content, built and developed by OpenAIRE and CERN (https://zenodo.org/)
- **Figshare** - repository where researchers can make all of their research outputs available in a citable, shareable and discoverable manner (https://figshare.com/)
- **Dryad Data Platform** - general-purpose open source platform for data publication and digital preservation of a wide diversity of data types, in underlaying repository repository with which it is integrated (https://datadryad.org/stash)
- **EUDAT** - Collaborative Data Infrastructure (or EUDAT CDI) is one of the largest infrastructures of integrated data services and resources supporting research in Europe (https://www.eudat.eu/)
- **Dataverse** – available as an open source web application to share, preserve, cite, explore, and analyze research data, developed by the Institute for Quantitative Social Science (IQSS) in collaboration with the Harvard University (https://dataverse.org/)
- **ELIXIR** – coordinates and develops life science resources across Europe so that researchers can more easily find, analyse and share data, exchange expertise, and implement best practices (https://elixir-europe.org/platforms/data/elixir-deposition-databases)
- **PAZy** - The Plastics-Active Enzymes Database - database that lists exclusively biochemically characterized plastic-active enzymes (https://pazy.eu/doku.php)
- **GenBank database –** a part of the International Nucleotide Sequence Database Collaboration database, designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information (https://ftp.ncbi.nih.gov/genbank/)

APPENDIX I. SAMPLE METRIC QUESTIONS FOR ASSESSING SECURITY AWARENESS AND TRAINING EFFORTS

Critical Element
Have employees received adequate training to fulfill their security responsibilities?

Subordinate Question
Are employee training and professional development documented and monitored?

Metric
The percentage of employees with significant security responsibilities who have received specialized training.

Purpose
To gauge the level of expertise among designated security roles and security responsibilities for specific systems within the agency.

Implementation Evidence

- Are significant security responsibilities defined, with qualifications criteria, and documented?
- Are records kept of which employees have specialized security responsibilities?
- How many employees in your agency (or agency component, as applicable) have significant security responsibilities?
- Are training records maintained? (Training records indicate the training that specific employees have received.)
- Do training plans state that specialized training is necessary?
- How many of those with significant security responsibilities have received the required training stated in their training plan?
- If all personnel have not received training, state all reasons that apply:
    - Insufficient funding
    - Insufficient time
    - Courses unavailable
    - Employee has not registered
    - Other (specify)

Frequency
Annually, at a minimum.

Formula
Number of employees with significant security responsibilities who have received required training (Question f) / Number of employees with significant security responsibilities (Question c).

Data Source
Employee training records or database; course completion certificates.

Indicators

The target for this measure is 100 percent. If security personnel are not given appropriate training, an organization may not be equipped to combat the latest threats and vulnerabilities. Specific security control options and tools are rapidly changing and evolving. Continued training enforces the availability of necessary security information.

APPENDIX J. INFORMED CONSENT FORM EXAMPLE

I confirm that I have read and understood the information sheet for the above study and have had the opportunity to ask questions. I am informed about the nature and objectives of the research, including the details and procedures necessary for conducting the study. I understand that my participation is voluntary and I am free to withdraw at any time without giving a reason and without any consequence. I am informed that results of the Study can be used for teaching and education, scientific meetings presentations, and publishing in scientific journals and doctoral dissertations. I am informed that my withdrawal will not affect my medical care, medical care of my child or child whose legal guardian I am. In case I have more questions, or wish to withdraw my participation, I can contact the members of the research team.

I confirm that the significance of this Consent has been explained to me in a comprehensible way.

I voluntarily agree to participate in the aforementioned Study and agree:

1. with the participation of me/my child/child whose legal guardian I am in this research study for the purposes stated in the Information sheet for this study

2. to receive a signed and dated copy of this Consent form for my archive. By signing this form, I am not waiving any legal rights

3. that you collect the socio-demographic data (age, gender, marital status, etc.) of me/my child/child whose legal guardian I am

4. that you perform complete physical examination of me/my child/ child whose legal guardian I am (measuring body height, weight and head circumference, examination of the heart beating, tapping stomach and more)

5. to provide information about the health and psychiatric history of me/my child/child whose legal guardian I am

6. to provide information about the psychiatric symptoms of me/my child/ child whose legal guardian I am

7. that you perform tests for evaluation of cognitive functions of me/my child/child whose legal guardian I am

8. that you analyze socio-emotional, speech and language abilities of me/my child/child whose legal guardian I am

9. that the research team can contact me again if they need further information and if more tests need to be done such as biochemical blood examination, ultrasound examination, electroencephalography, magnetic resonance imaging

10. that the results obtained in this study could be published in scientific journals, deposited in databases, presented in scientific conferences in the country and abroad and used for teaching and training, as well as for the preparation of doctoral dissertations

11. that principal investigator and researchers involved in the study have access to my/my child's/child's whose legal guardian I am data collected during the study. Any information related to me will be kept confidential. Only anonymized data will be used for publications.

12. that my/my child's/child's whose legal guardian I am biological and medical samples, data and results obtained in this study are shared with other centers in the world for research purposes. Data and results will not include, without my explicit permission, names, photos or other documents that can identify me/my child/child whose legal guardian I am.

Name and surname of participant/parent/legal guardian:

Signature: _____

Name and surname_____

Child's name and surname _____

Signature of child (optional) _____

Telephone number: _____

Date:_____ Place: _____

Name and surname of researcher:

Signature: _____

Name and surname _____

Date:_____ Place: _____

REFERENCES:

1. OpenAIRE Webinar: Horizon Europe Open Science requirements in practice, https://zenodo.org/record/6641829#.Y9arQq3MJpk
2. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016)
3. Practical Guide to the International Alignment of Research Data Management (Extended Edition), Science Europe, January 2021
4. UK Data Service. Store your data. URL: https://www.ukdataservice.ac.uk/manage-data/store.aspx
5. OpenAIRE. Amnesia. URL: https://amnesia.openaire.eu/
6. Guide for data documentation, 2020. URL: https://www.helsinki.fi/en/research/guide-for-data-documentation
7. Horizon Europe (HORIZON), Programme Guide, Version 1.4, 17 December 2021
8. UK Data Service. Anonymisation. URL: https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation/quantitative.aspx
9. Plan S Rights Retention Strategy, https://www.coalition-s.org/rights-retention-strategy/
10. Data Management Plan - CORDIS, Project 101046758 — EcoPlastiC
11. UK Data Service. File formats and software. URL: https://www.ukdataservice.ac.uk/manage-data/format/file-formats.aspx
12. Jisc. Research Data Management Toolkit. URL: https://rdmtoolkit.jisc.ac.uk/manage-store-and-preserve/security/
13. Digital Curation Centre. URL: https://www.dcc.ac.uk/resources/data-management-plans/guidance-examples
14. AGA — Annotated Model Grant Agreement, EU Funding Programmes 2021-2027: V0.2 DRAFT– 30.11.2021
15. Practical Guide to the International Alignment of Research Data Management (Extended Edition), Science Europe, January 2021
16. Plan S and the sharing of Author Accepted Manuscripts without embargo and with a public copyright licence (Rights Retention Strategy), https://www.coalition-s.org/wp-content/uploads/2020/07/Letter_to_publishers_Rights_Retention_Strategy_15July2020.pdf
17. Research Data Management Plan: guidance and resources, Dublin City University, February 2022
18. RDA Metadata Standards Directory, http://rd-alliance.github.io/metadata-directory/
19. FAIR principles, https://force11.org/info/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-0/
20. Istraživački podaci - što s njima? : priručnik o upravljanju istraživačkim podacima https://repozitorij.srce.unizg.hr/islandora/object/srce%3A327/datastream/FILE0/download
21. Angus Whyte (DCC) and Andrew Wilson (ANDS), How to Appraise & Select Research Data for Curation, Digital Curation Centre, Australian National Data Service, 2010
22. Horizon Europe – (2021. — 2027.) https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

23. University of Edinburgh, Research Data Service: Our definitions. https://www.ed.ac.uk/information-services/research-support/research-data-service/after/data-repository/definitions
24. Utrecht University Guides. Storing and preserving data. URL: https://www.uu.nl/en/research/research-data-management/guides/storing-and-preserving-data
25. Rehberger, Dean; Coates, Brendan. File naming in the digital age. URL: http://ohda.matrix.msu.edu/2012/08/file-naming-in-the-digital-age/
26. Cornell University Libraray. Recommended File Formats. URL: https://guides.library.cornell.edu/ecommons/formats
27. Corti, L., Van den Eynden, V., Bishop, L. & Woollard, M. Managing and Sharing Research Data: A Guide to Good Practice. SAGE Publications (2020).
28. Piwowar HA, Vision TJ. 2013. Data reuse and the open data citation advantage. PeerJ 1https://doi.org/10.7717/peerj.175
29. Krishna, Vasmi. How to tag files in windows for easy retrieval, 2018. URL: https://www.maketecheasier.com/tag-files-in-windows/
30. Swiss National Science Foundation. Data Management Plan (DMP) — Guidelines for researchers. URL: http://www.snf.ch/en/theSNSF/research-policies/open_research_data/Pages/data-management-plan-dmp-guidelines-for-researchers.aspx
31. The European Union Agency for Cybersecurity. Pseudonymization techniques and best practices. URL: file:///C:/Users/Administrator/Downloads/Guidelines%20on%20shaping%20technology%20according%20to%20GDPR%20provisions.pdf
32. UK Data Service. Data management planning. URL: https://www.ukdataservice.ac.uk/manage-data/plan/planning.aspx
33. Stanford University. Sharing sensitive data. URL: https://library.stanford.edu/research/data-management-services/share-and-preserve-research-data/sharing-sensitive-data
34. DePaul University Library. Research Data Management (A How-to Guide). URL: https://libguides.depaul.edu/c.php?g=620925&p=4324498
35. Digital Curation Centre. URL: https://www.dcc.ac.uk/resources/data-management-plans/guidance-examples
36. Educopia Institute. Preservation and Curation of ETD Research Data and Complex Digital Objects. URL: https://educopia.org/wp-content/uploads/2018/04/etdplus_storage_guidancebrief.pdf
37. Force11 — The FAIR Data Principles. https://www.force11.org/group/fairgroup/fairprinciples
38. GoFAIR — FAIR Principles https://www.go-fair.org/fair-principles/
39. Guide to writing "readme" style metadata. URL: https://data.research.cornell.edu/content/readmehttps://learn.canvas.net/courses/2719/pages/exercise-2-readme-file-faculty-follow-up?module_item_id=241426
40. Data Archiving and Networked Services. Preferred formats. URL: https://dans.knaw.nl/en/about/services/easy/information-about-depositing-data/DANSpreferredformatsUK.pdf
41. UK Data Service. File formats and software. URL: https://www.ukdataservice.ac.uk/manage-data/format/file-formats.aspx