OpenAIRE Guide to storing sensitive data

THE CONTEXT

What is the best way of managing access to sensitive data? This is not a straightforward question as it involves ethical, legal and technical issues to be tackled. This guide will help you on your way to preserve your sensitive data safely. It explains the different types of sensitive data, how to prepare them for storage and the possible cost involved in the process.

WHAT IS SENSITIVE DATA?

Sensitive data is data that must be protected against unwanted disclosure. Access to sensitive data should be safeguarded. Protection of sensitive data may be required for legal or ethical reasons, for issues pertaining to personal privacy, or for proprietary considerations.

The EU has strong regulations regarding personal (e.g. General Data Protection Regulation - GDPR) and non-personal data. In Horizon 2020 an Ethical review is required when applying for a grant.[1]

Examples of sensitive data are:

- Personal data: identifiers such as names or identification numbers, physical, physiological, genetic, mental, economic, cultural or social characteristics, it also includes location data from GPS or mobile phones
- Confidential data: trade secrets, investigations,data protected by intellectual property rights
- Security: passwords, financial information, national safety, military information...
- Combination of different datasets that can be combined into sensitive or personal data
- Biological data: endangered (plant or animal) species, where their survival is dependent on the protection of their location data (biodiversity community)[1] ● Personal and sensitive metadata[2]

[1] EUDAT Sensitive Data Group Working Paper: How can e-infrastructures deal with the sensitive data challenge (Working Paper), 2017, p. 3.

[2] EUDAT Sensitive Data Group Working Paper: How can e-infrastructures deal with the sensitive data challenge (Working Paper), 2017, p. 4.

When handling and dealing with sensitive data, keep in mind that special attention should be given to collecting, processing, handling and storing data throughout the research process. In particular research data that contains personal data with which a living person can, directly or indirectly, be identified has to be handled with care. This concerns both textual data and image and sound data. Examples of direct data are someone's name and address, but it could also be [2] a photo or an interview. An indirect fact is, for example, someone's employer. For personal data fully informed consent should be given for collecting, processing and storing data.

How to prepare sensitive data for storage and sharing?

When storing sensitive data, the first concern is to find a good security strategy for your type of data.

Sensitive data can still meet the requirements of the FAIR Data Principles (findability, accessibility, interoperability, and reusability) and be processed in a way that the needed protection is guaranteed also in the future.

Anonymization

- Anonymization irreversibly destroys any way of identifying the data subject.

  Personal data that has been rendered anonymous in such a way that the individual is not or no longer identifiable is no longer considered personal data. For data to be truly anonymised, the anonymisation must be irreversible.

  OpenAIRE provides researchers with a tool to anonymise data: Amnesia. The guide for which you can find here.

Pseudonymization

- Pseudonymization substitutes the identity of the data subject in such a way that additional information is required to re-identify the data subject. The pseudonym allows tracking back of data to its origins, which distinguishes pseudonymization from anonymization, where all person-related data that could allow backtracking has been purged. Pseudonymised data are still legally considered as sensitive data because the data can be linked back to a person, but it's considered as a secure approach since personal identifiers are stored somewhere else.

Encryption

- Encryption is a very generic term and there are many ways to encrypt data. The key to a good encryption strategy is using strong encryption and proper key management. Encrypt sensitive data before it is shared. Encryption will make your data totally unintelligible to those who may try to access it which might reduce re-usability.

In case none of these options are available for your dataset, data should not be made open and be archived under a closed license in a Trustworthy Repository. You can however publish a description (i.e.public metadata) of your data without making the data itself openly accessible, which enables you to place conditions around access to the data.

There are other ways to share sensitive data too:

- Information professional:is someone who collects, records, organises, stores, preserves, retrieves, and disseminates printed or digital information.
- Data access committees: A committee that reviews and authorizes applications for data access and use.
- Safe havens: provide access to data and services to enable research while protecting the confidentiality of the data.
- institutional data archives/vault: safe, private, store of data that is only accessible by the data creator or their representative.

STORING SENSITIVE DATA

To make sure your data is archived securely you should look for a Trustworthy Repository.

If possible choose a certified repository that matches the needs of your data regarding file formats, access and licensing. Look for a repository that provides a persistent and globally unique identifier, guidance on data citation and clear information on costs (if any). Re3data.org can give you information on what data repositories provide in terms of information, access, licensing, persistent identifiers, policy and certificates and standards.

In case you are archiving sensitive data you should always go for a restricted license or closed access. You should also keep in mind that it is not always necessary to keep all your data. What you need to store will likely depend on criteria such as their uniqueness, long-term value, reuse potential and the necessary to validate results in publications.

When archiving your research data keep in mind to describe it with metadata so it can be findable for other researchers. Make sure the repository you've selected supports the metadata standards you are using for your data. You can find further information on metadata for research data here: What is metadata.

Remember that it can also be useful to store consent forms to the data, with restricted or closed access.

COSTS

The cost related to open access to research data are eligible as part of the Horizon 2020 grant - refer to your Grant Agreement conditions.[3]

For further information, see our guide on cost in research data management.

How can OpenAIRE help

Pages

How to find a trustworthy repository

About GDPR and the research process

Data Protection and Ethics course

Tools

Amnesia is a data anonymization tool.

Guides

[3] Copied from Sarah/Creating Data: https://docs.google.com/document/d/1BPNl70iiQ2yQX5Vge2nXsLe8px6gajbfELJdtEDHw0/edit#heading=h.zc5aduauvwi3

Amnesia - How to use the data anonymization tool

How to comply to the H2020 mandate for data

RDM costs in H2020

Factsheets

What is personal data