Guide to Secure Data Storage

A PD-Net Study Process Template

Data collected in the context of PD-Net should be protected from loss, corruption, and unauthorized access. Of particular importance is the confidentiality of **personally identifiable information (PII)**.

# Secure Data Storage – Core Concepts

The concept of personally identifiable information (PII) expresses the fact that we want to prevent collected information to be *linked* to an individual. Linking requires that a person can be distinguished – with more or less certainty – from all other persons. PII is what allows us to perform this linking. Examples of PII are, e.g., the first and last name; a home or other physical address including street name and name of city or town; an email address; and a telephone number. Also a seemingly random identifier, such as an IP address or a **Bluetooth MAC** address, can become PII, if these can in turn be linked to any other PII such as a physical address.[[1]](#footnote-1)

When storing data related to user studies and experiments in PD-Net, we thus need to either protect access to all and any data collected from participants, or alternatively ensure that this data remains *unlinkable* to the participants, by separating and safeguarding all PII. Note, however, that the concept of privacy is much broader than just ensuring the secure storage of PII. While this document describes only the secure data storage principles adopted in PD-Net, please refer to the PD-Net Ethics Primer for a comprehensive overview of privacy requirements in PD-Net.

## Pseudonymous Identifiers

A typical approach in user studies is to use *pseudonymous* *identifiers* for all participants, and to store the data that links the participants’ identity to those identifiers separately from the collected data (also called “coded identification”). The researcher maintains a list that contains a link between the subject’s name and a random code number or pseudonym, and then uses only the code number/pseudonym to mark the data and responses from that subject. The list that contains the coded link should be kept secure, and separate from the data and responses – if stored in electronic form it must be encrypted, if on paper it should be kept in a safe place (e.g., locked office). Obviously, care must be taken so that the remaining data does not involuntarily contain other PIIs, such as the above-mentioned IP addresses or Bluetooth MAC addresses, or demographic information from a small group that implicitly allows identification. Also, some study data such as video and audio recordings might inherently allow for the identification of a participant, and thus may need to be protected in their entirety.

If numeric identifiers such as IP addresses or Bluetooth MAC addresses are to be stored in system logs, these should ideally be anonymized before storage, so that the logs themselves can be stored without additional security precautions (i.e., without having any PII embedded). Note that simple hashing of such identifiers does *not* provide adequate anonymization of PII, as long as the set of potential matches can easily be enumerated (e.g., a set of students in a class). If later re-identification of such values is needed, use an encrypted hash (as “message authentication code”, MAC) and store the key with other PII data in a separate, encrypted file.

## Anonymization

 If you want to fully anonymize collected identifiers such as phone numbers or Bluetooth MAC addresses, simply use a one-time key for computing an encrypted hash, then destroy the key. You might also want to consider hashing only part of the identifier, e.g., the last 6 digits of a Bluetooth MAC address, but leave the first 6 digits in the clear. This allows, e.g., to identify devices and classes of devices. See also [1] for a more detailed discussion of anonymization techniques of common PII. Note that even where only anonymised data are used, *adequate* security for storage and handling of such data must be ensured [6].

## Adequate Protection

The level of protection for all stored data should be commensurate to the expected *impact level* of any inadvertent disclosure of the data. The recommendations in this brief describe the baseline for secure data storage to be adopted in PD-Net – refer to the risk assessment portion (section 4) of the *PD-Net Ethics Worksheet* for a particular user study in order to see whether your study might warrant higher levels of protection. In general, AES encryption should be employed with key strength of at least 64 bits, which corresponds to passwords of at least 10 characters using a mix of uppercase and lowercase, numbers, and special characters. See the useful software links at the end of this brief for links to password generators. If sensitive data (in the sense of directive 1995/46/EC) is recorded, higher protection levels (up to 128 bits) might be necessary. Data may be individually encrypted using application-specific encryption formats (e.g., Word documents) and archive storage formats (e.g., ZIP files), though a transparent directory-based or partition-based disk-encryption approach may offer both a higher level of protection and simplification of access.

# Secure Data Processing Principles

The list below outlines the core principles of secure data storage in PD-Net.[[2]](#footnote-2) They need to be taken into account for *all* data collections in PD-Net, i.e., both for records of user studies and interviews, as well as for system logs that record PII (e.g., Bluetooth MAC addresses).

## General Principles

1. Code data as early as possible, i.e., replace all PII with anonymous identifiers.
2. Keep the coding key[[3]](#footnote-3) separate (in a physically separate space or in a separate electronic file) from the data.
3. Work with de-identified data at all times, unless this is not possible for your work.
4. Never write down passwords. Use a password-storage application if needed (see the Useful Software links below), together with a strong Master Password.

## Electronic Data Safekeeping

1. Store all files containing PII on encrypted drives, or use encrypted files. Ensure that no temporary unencrypted copies are left (e.g., trash).
2. Ensure that you always have a backup of your (encrypted) files in another location, e.g., on another server of your university, in order to avoid data loss.
3. If you need to remotely access data, always use a VPN connection or an encrypted Remote Desktop Session.

## Physical Storage

1. Avoid hard copy media (e.g., video tapes, DVDs, printed documents) for storing PII as much as possible.
2. Store any such media in an institutional environment with restricted access and lockup capability (e.g., in your office in a desk that is locked at all times).
3. If the contents of the media is crucial to your research, ensure that a backup copy is (equally safely) stored in another physical location on campus to avoid data loss from, e.g., fire.
4. Only take such information off-site if absolutely necessary. If you must do so, take all reasonable security precautions consistent with protecting a high-value asset.

## Data Processing

1. Do not store or disclose personally identifiable or confidential data other than as necessary for your research
2. Keep an accurate and up-to-date log, detailing your use of personally identifiable and/or confidential data and the specific security and privacy protection measures that you apply.
3. Immediately report privacy concerns (like possible data loss) to the PI and/or your local ERB contact.

## Data Retention and Deletion

1. Ensure that records are retained only as long as is required to accomplish research purposes and satisfy legal and policy retention requirements. Note that all PII collected in PD-Net must be deleted at the end of the project, though local retention requirements might take precedence.
2. Ensure the secure destruction of all personally identifiable or confidential information at the end of applicable retention periods. Documents must be shredded, video tapes re-recorded, DVDs and CDs securely deposed, and electronic files must be securely deleted using repeated overwriting.
3. If you made additional copies of your data for backup purposes, (securely) delete all on-line copies (e.g., on servers). If additional off-line copies remain due to institutional backup procedures (e.g., on tape or WORM media), you *must* ensure that all *access keys* are securely deleted (see points 4 and 15 above).

# Useful Software

Below you will find some recommendations for secure data storage and processing software. These may help in implementing some of the guidelines mentioned in this document.

## Encryption Tools

**Truecrypt** (MacOSX, Linux, Windows): A free (i.e., without cost) on-the-fly encryption (OTFE) tool. Can encrypt entire partitions transparently, or create new virtual disks that can be mounted as if they were an external disk. Cross-platform.

**FileVault**, **DiskUtility** (MacOSX): Both tools are part of recent versions of MacOSX. FileVault allows users to encrypt the entire home directory, DiskUtility can be used to encrypt individual directories.

For individual files, a simple password-protected zip-file may also be sufficient. The free cross-platform compression utility **7-zip** offers AES-based file encryption. This might also be sufficient if sharing files with other project members via email. Obviously, the password must not be shared via email or instant messaging apps, but only in person or via a phone call.

A *Password Safe* application such as **KeePass** (Windows) or the cross-platform equivalent **KeePassX** (Windows, Linux, MacOSX) can be used to safely store a large number of passwords. Many other (free and non-free) alternatives exist. These applications also allow the creation of random passwords of arbitrary length using a built-in *Password Generator*, which greatly improves the entropy of the used passwords. See [1],[4] for hints on how to come up with (and remember!) a good master password.

If you regularly exchange such files via email, you should consider the use of **encrypted email**. Free PGP-based extensions are available for most modern email programs (e.g., the Enigma plug-in for Mozilla Thunderbird). Note that the recipient must use the same encryption system (either PGP/GnuPG or S/MIME based).

## Secure Deletion Tools

MacOSX already comes with a secure delete option with its standard Trash. “**Secure Empty Trash**” is an option that allows one to securely wipe the entire trash (File > Secure Empty Trash). The system will then perform a 35-pass over the files contained within the trash.

Linux systems (e.g., Ubuntu) can use the built-in **shred** command to securely overwrite a file. You might also want to use the **scrub** command to wipe *unused* space periodically. Also, the open source “**Wipe**” utility (<http://wipe.sourceforge.net/>) can be used. See also [5] for further information.

## Bibliography

1. Aad, Imad, and Valtteri Niemi, “NRC Data Collection and the Privacy by Design Principles”. Proceedings of the 1st Workshop Sensing for App Phones (PhoneSense 2010), at SenSys 2010, Zurich, November 2010
2. Data Security Standards for Personally Identifiable and Other Confidential Data in Research. University of Toronto, Research Ethics Board (REB), April 2010
3. “Lockdown Password Guidelines.” Lockdown.co.uk - The Home Computer Security Centre. July2009. Available from [www.lockdown.co.uk/?pg=password\_guide](http://www.lockdown.co.uk/?pg=password_guide)
4. “Microsoft Advisory: Create Strong Passwords”. Microsoft Online Safety Blog, Microsoft. Available from [www.microsoft.com/protect/fraud/passwords/create.aspx](http://www.microsoft.com/protect/fraud/passwords/create.aspx)
5. Peter Gutmann, „Secure Deletion of Data from Magnetic and Solid-State Memory”, Sixth Usenix Security Symposium, 1996. Extended version at [www.cs.auckland.ac.nz/~pgut001/pubs/secure\_del.html](http://www.cs.auckland.ac.nz/~pgut001/pubs/secure_del.html)
6. Eléonore Pauwels: “Ethics for Researchers. Facilitating Research Excellence in FP7”. See ftp://ftp.cordis.europa.eu/pub/fp7/docs/ethics-for-researchers.pdf
1. For example, many legislators today consider IP addresses as PII, since providers (e.g., ISPs) can typically associate an IP address with a physical address for any point in time, even if these are being repeatedly changed (e.g., for DSL access). [↑](#footnote-ref-1)
2. These principles are modeled after [2] [↑](#footnote-ref-2)
3. In this case, the term „key“ refers to the index that links the anonymous identifier to PII of the individual study subject, not to an „encryption key“ per se. However, if it is impossible to remove PII from the actual data collection, the entirety of the data must be encrypted with a sufficiently long key and the key must be kept separately as described in principle 2. [↑](#footnote-ref-3)