good brother

Network on Privacy-Aware
Audio- and Video-Based Applications
for Active and Assisted Living

Data Management Plan

Date: 26 January 2023

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

Funded by the Horizon 2020 Framework Programme
of the European Union

Table of contents

1. Data summary

1.1. Introduction to GoodBrother

The GoodBrother COST Action (CA19121) aims to enhance the quality of life for older, impaired, and frail individuals by developing privacy-aware solutions for Active and Assisted Living (AAL). It focuses on raising awareness of the ethical, legal, social, and privacy concerns associated with audio- and video-based monitoring in private spaces.

By fostering an interdisciplinary community of researchers and industry partners from fields such as computing, engineering, healthcare, law, and sociology, GoodBrother seeks to stimulate new research and innovation. This collaboration includes stakeholders like users, policymakers, and public services, aiming to mitigate the "Big Brother" perception of continuous monitoring. The goal is to increase user acceptance, promote the adoption of these solutions, and expand their market reach.

GoodBrother is structured into five Working Groups (WGs):

- WG1: Social Responsibility – Addresses ethical, legal, social, data protection, and privacy issues.
- WG2: Privacy-by-Design – Focuses on integrating privacy considerations into audio and video data processing.
- WG3: Audio- and Video-Based AAL Applications – Develops applications utilising audio and video technologies for AAL.
- WG4: Curated Repository of Software and Data – Maintains a repository of relevant software and datasets.
- WG5: Dissemination and Exploitation – Promotes the dissemination and practical application of research findings.

1.2. Purpose of the data collection/generation

The GoodBrother COST Action collects and generates data to explore privacy-aware solutions and examine ethical, legal, and privacy challenges associated with audio- and video-based monitoring technologies. These efforts aim to balance functionality, user trust, and privacy by addressing societal, legal, and technological questions. The data could support various disciplines and applications, contributing to the development of solutions that improve quality of life and encourage innovation in AAL systems. The following outlines the potential purposes for data collection across different research areas:

<u>Understanding public perceptions and acceptance</u>

Social science research might examine how individuals and communities perceive video- and audio-based technologies in AAL. Potential areas of exploration include:

- Privacy perceptions: Researchers might investigate how privacy concerns are influenced by the type of technology, its context of use (e.g., home, care homes, public spaces), and the nature of the data collected (e.g., audio, video, metadata). This could include studying individuals' concerns about third-party misuse of their data.

- Trust in technologies: Studies might assess levels of trust in lifelogging technologies, particularly in sensitive contexts such as healthcare. Factors influencing trust, such as transparency in data handling or the perceived reliability of systems, could be explored.
- Acceptance of AI-based solutions: Research might focus on understanding attitudes towards artificial intelligence (AI) in medical and care-related technologies. Misconceptions about AI and its potential impact on individuals' quality of life may be key areas of interest.

Data collected in this area might support multidisciplinary analyses of user acceptance, accounting for cultural, societal, and gender-based factors.

Addressing legal and ethical challenges

Legal studies could investigate the implications of video- and audio-based monitoring systems in various environments, including public and private spaces. Possible data sources and topics include:

- Case law and precedents: Researchers might analyse legal cases and regulatory frameworks to understand the legal requirements and constraints for using monitoring technologies.
- Sensitive personal data: Some datasets could include information such as health status or religious beliefs, helping to explore the ethical dimensions of data use in monitoring contexts.
- Context-specific issues: Studies might look at the ethical trade-offs between privacy and benefits in environments like care homes or shared spaces.

These investigations might inform how privacy laws and regulations could evolve to better protect vulnerable populations, such as older adults.

Evaluating health and wellbeing outcomes

Health-focused research might evaluate the impact of AAL technologies on older adults and their care networks. Potential goals include:

- Health and behavioural impact: Researchers might collect data to assess outcomes like mobility, fall detection, or stress management in older adults using these technologies.
- End-user experiences: Data could be gathered to explore how users, including caregivers and healthcare professionals, perceive and interact with these systems.
- Implementation readiness: Studies might assess whether audio- and video-based monitoring systems are feasible for real-world deployment and identify potential challenges.

Such data could help ensure that AAL technologies are well-suited to the needs of their intended users while promoting independence, safety, and quality of care.

Technological research and development

Technological studies might focus on improving AAL systems by collecting data to support development and validation. Potential purposes include:

- Training machine learning models: Data, including video and audio recordings, might be used to train algorithms for tasks like fall detection, behaviour analysis, or privacy preservation.
- Developing systems: Researchers could explore how to create systems that are effective while respecting user privacy.
- Dataset expansion: Where existing datasets are insufficient, new data might be collected. This could include RGB images, thermal images, audio files, or smart home sensor data.

The insights gained might contribute to creating systems that are not only functional but also ethically acceptable to end users.

1.3. Types and formats of data generated/collected

The GoodBrother COST Action might generate and collect various types of data related to both audio- and video-based technologies. These data could support research across social sciences, legal studies, health-related investigations, and technological development. The types and formats of data that might be used or produced are detailed below.

Social science data

Social science research might collect qualitative and quantitative data to explore user needs, perceptions, and acceptance of audio- and video-based technologies in AAL.

- Types of data:
    - Qualitative data: Transcripts from interviews and focus groups with potential end users (e.g., older adults, caregivers, healthcare professionals).
    - Quantitative data: Responses to surveys or questionnaires, including demographic data and opinions on privacy and technology acceptance.
    - Multimedia data: Audio and video recordings of discussions or feedback sessions.
- Formats:
    - Textual data: .txt, .csv, .xls for raw survey data; .sps (SPSS) and .R (R programming) for statistical analysis.
    - Audio/video data: Common formats such as .wav, .mp3 for audio, and .mp4 for video.
    - Transcriptions: .docx or .txt for textual versions of recorded sessions.

Identifiable information, such as names, voices, or images, might be collected initially but would be anonymised or pseudonymised before analysis to comply with data protection standards.

Legal data

Legal studies might rely on secondary data from publicly available legal documents and academic sources, examining issues related to the use of audio- and video-based monitoring technologies.

- Types of data:
    - Case law and legal decisions involving audio and video monitoring.
    - Statutes, regulations, and guidelines on privacy and data protection.
    - Secondary sources, such as academic articles and reports.
- Formats:
    - Text documents (.txt, .docx, .pdf) for legal texts and analyses.
    - References in bibliographic formats compatible with tools like EndNote or Zotero.

This data might include sensitive personal information, such as individuals' names in court cases. If collected, it would likely be anonymised in accordance with GDPR.

Health and wellbeing data

Health-related research might evaluate how AAL technologies affect the health and wellbeing of older adults and their care networks.

- Types of data:
    - Metrics related to movement, fall detection, or emotional state monitoring.
    - Feedback from potential users, including caregivers, healthcare providers, and older adults.
    - Recordings of participants interacting with audio- and video-based AAL systems.
- Formats:
    - Quantitative data: Spreadsheet files (.csv, .xls) for numerical data.
    - Qualitative data: Transcripts of interviews or focus groups (.txt, .docx).
    - Multimedia data: Video (.mp4) and audio (.wav, .mp3) files recorded during usability testing.

Such data might inform the development of systems tailored to meet user needs, while maintaining privacy.

Technological data

Technological projects might focus on datasets to train and validate machine learning algorithms for AAL applications.

- Types of data:
    - Audio data: Speech and sound recordings for activities such as interaction recognition or emotional state analysis.
    - Video and image data: RGB, depth, infrared, or thermal recordings of monitored environments.
    - Metadata: Contextual information, such as timestamps, activity labels, or sensor readings.
- Formats:
    - Audio data: .wav, .mp3 for speech and environmental sounds.
    - Video data: .mp4, .avi for standard video; specialised formats for depth or infrared data; .png or .jpg for images.
    - Metadata: Structured formats such as .csv or .json.

In cases where public datasets are insufficient, researchers might collect new data in controlled environments, such as smart homes or care facilities. These datasets would likely be handled under strict data protection protocols.

1.4. Re-use of data

The GoodBrother COST Action might involve the re-use of existing datasets, depending on the specific requirements of each study and the availability of relevant data. Any re-use would comply with GDPR and other legal requirements, ensuring alignment with ethical and technical standards. Researchers might use publicly available datasets or previously collected data where appropriate safeguards are in place, supporting efficient and responsible research practices. The following considerations outline how re-use might be approached:

Conditions for re-use

Re-use of data would need to adhere to applicable legal, ethical, and technical requirements:

- Publicly available datasets: Researchers might use datasets that are publicly accessible, such as those in academic repositories or datasets shared under open licences (e.g., Creative Commons). Adherence to the terms and conditions of these datasets, including licensing and citation requirements, is essential.
- GDPR compliance: The General Data Protection Regulation (GDPR) establishes conditions for further processing of personal data:

- If data were collected based on consent (Article 6(1)(a)) or legal obligations (Article 6(1)(c)), additional processing might require new consent or a new legal basis.
- If data were collected under legitimate interest (Article 6(1)(f)), contractual obligations (Article 6(1)(b)), or vital interests (Article 6(1)(d)), re-use might only proceed after assessing compatibility with the original purpose.
- Data used for statistical or scientific research purposes might be exempt from the compatibility test, provided safeguards such as pseudonymisation are in place (Articles 5(1)(b) and 89(1)).

Factors to consider

The potential re-use of data would involve evaluating:

- Purpose alignment: Whether the original purpose of data collection aligns with the intended use in the current project.
- Nature of the data: The type of data, including whether it involves sensitive information such as health, biometric, or personal details.
- Context of collection: The relationship between the original data collector and the subjects, and whether that context allows for re-use.
- Potential impact on individuals: How further processing might affect data subjects, including risks to privacy or personal wellbeing.
- Safeguards: Measures such as encryption, pseudonymisation, and access controls to protect data during re-use.

Anticipated scenarios for re-use

Re-use might occur in several contexts:

- Benchmark datasets: Technological projects might leverage publicly available datasets to train and validate machine learning models, particularly for tasks involving activity recognition or behaviour analysis.
- Survey or study data: Previous surveys or studies on related topics might provide useful insights for comparison or supplementary analysis, provided licensing conditions are met.
- Secondary legal data: Legal studies might re-analyse case law, statutory guidelines, or academic articles to explore privacy or ethical concerns.

Re-use in these cases would respect all relevant licensing agreements, attribution requirements, and ethical standards.

Potential limitations

While re-use might reduce the need for new data collection, certain challenges could arise:

- Data licensing: Restrictions on access or re-use might limit the scope of certain datasets.
- Data format: Variability in formats or structures might require additional efforts to standardise datasets for use.
- Privacy concerns: Sensitive data might not always be available for re-use due to ethical considerations or data protection regulations.

1.5. Origin of the data

The data used in GoodBrother projects might come from various sources, including primary data collected during the project, publicly available datasets, or data shared by external entities. The origin of the data would depend on the research area, with recruitment and collection processes prioritising transparency, informed consent, and adherence to GDPR requirements. By sourcing data responsibly, GoodBrother aims to generate high-quality insights to support its interdisciplinary goals. The following outlines potential data origins across different research domains.

Social science research

Social science projects might collect primary data through direct engagement with participants to understand user needs and perceptions regarding audio- and video-based technologies in AAL. This could include:

- Empirical studies: Data might be collected through interviews, focus groups, or surveys with older adults, caregivers, healthcare professionals, or family members.
- Participant selection: Recruitment might involve applying specific inclusion criteria (e.g., age, health status, communication abilities) to ensure participants can provide relevant information. Recruitment might also be delegated to market research institutions, which operate under strict data protection and privacy standards.
- Informed consent: All participants would be required to provide informed consent, ensuring compliance with GDPR Article 13. Data collection processes would prioritise transparency and respect for participants' rights.

These studies might yield diverse perspectives, enabling a comprehensive analysis of user acceptance, privacy concerns, and ethical considerations.

Legal studies

Legal research might use secondary data sourced from publicly available legal documents, including:

- Case law and precedents: Legal cases involving privacy, data protection, and the use of monitoring technologies might be analysed.
- Regulations and guidelines: Statutes, legal frameworks, and data protection authorities' guidance could provide additional insights.

- Academic and policy resources: Secondary sources, such as academic articles or policy reports, might also be utilised.

These data sources are typically accessible online or through legal databases, with GDPR compliance ensuring the appropriate use of any personal data.

Health and wellbeing research

Health-related projects might collect primary data from participants and their care networks to assess the impact of AAL technologies. This data could originate from:

- Interviews and surveys: Caregivers, healthcare professionals, and older adults might provide qualitative or quantitative data on their experiences and perceptions of AAL systems.
- Usability testing: Data might be gathered from observing participants interacting with audio- and video-based technologies in controlled or real-world settings.
- Collaborations: External organisations, such as healthcare providers or research institutions, might share anonymised data to support specific analyses.

Recruitment strategies might involve targeted outreach to specific user groups, ensuring that the data collected aligns with the project's research goals.

Technological projects

Technological projects might rely on a combination of existing datasets and newly collected data:

- Existing datasets: Publicly available datasets shared within the research community might serve as a starting point for training and validating algorithms.
- Newly generated data: Where existing datasets are insufficient, new data might be collected in controlled environments, such as smart homes or care facilities. This could include recordings (audio, video) and sensor data capturing activities, behaviours, or environmental conditions.
- External contributions: Collaborations with other research institutions or organisations might provide access to additional datasets, subject to licensing agreements and ethical guidelines.

Data collection in technological projects would follow strict privacy and security protocols to ensure compliance with ethical standards.

1.6. Expected size of the data

The size of the data generated or collected in GoodBrother projects might vary significantly depending on the type of research and the methods employed. While precise estimates are not currently available, data volumes are anticipated to range from small text-based datasets (kilobytes or megabytes) to large multimedia or sensor datasets (gigabytes or terabytes). Storage requirements would depend on the scope of specific studies, but careful planning and

adherence to data minimisation principles are expected to keep data volumes manageable within the project's infrastructure. The following outlines potential data sizes across different research areas.

Social science and legal research

- Social science data:
    - Survey and questionnaire datasets might depend on the number of participants and questions, typically ranging from a few megabytes (MB) to tens of megabytes per study.
    - Qualitative data, such as interview or focus group transcripts, might add a similar volume. However, multimedia recordings (audio or video) could require significantly more storage, with an hour-long video potentially requiring several gigabytes (GB).
- Legal data:
    - Secondary sources, such as case law or legal statutes, are primarily text-based and expected to remain small in size, ranging from kilobytes to a few megabytes per file.

These datasets are unlikely to pose challenges for storage or processing due to their relatively manageable sizes.

Health and wellbeing research

- Survey and feedback data: Numeric and qualitative data from health-related studies, such as caregiver surveys or usability tests, are expected to remain small (up to tens of megabytes).
- Audio and video recordings: Data collected during usability testing or observational studies might be larger, depending on the duration and quality of the recordings. High-definition videos could require several gigabytes per session.

Despite the variability, data volumes in this area are not expected to exceed the storage capacity available within the project's infrastructure.

Technological research

- Public datasets: Datasets sourced from public repositories might range in size, with many benchmark datasets for machine learning spanning hundreds of megabytes (MB) to a few gigabytes (GB).
- Newly collected data:
    - Data collected in controlled environments, such as smart homes, might include high-resolution video and audio recordings, infrared or thermal images, and sensor logs. These could collectively range from gigabytes to terabytes (TB), depending on the scale and duration of the experiments.

- Metadata, such as activity labels or timestamps, is likely to be small in size (kilobytes to megabytes per dataset) but integral to the overall dataset.

For very large datasets, specific storage solutions might be required to ensure availability and accessibility while complying with data protection standards.

1.7. Data utility

The data generated or collected within GoodBrother projects might provide valuable insights across multiple domains, contributing to societal benefits, technological advancements, and academic knowledge. By addressing user needs, ethical concerns, and technical challenges, this data could support the design of privacy-aware audio- and video-based technologies while promoting their acceptance and implementation. Furthermore, the data might align with the priorities outlined in the European Commission's *Communication on Digital Transformation of Health and Care in the Digital Single Market*. These priorities include empowering citizens with digital tools for feedback and person-centred care, promoting multi-disciplinary care teams, and fostering research on the impact of innovative technologies. The potential utility of this data includes the following:

Improving quality of life for older adults

- Promoting independence: Data might be used to develop systems that enable older adults to live independently for longer in their own homes, enhancing their autonomy and safety.
- Enhancing wellbeing: Insights into user needs and experiences might inform the design of technologies that improve health and emotional wellbeing, such as systems for fall detection, activity monitoring, and mental health support.
- Balancing privacy and functionality: Data might guide the development of solutions that address privacy concerns, fostering trust and acceptance among users.

Supporting ethical and legal frameworks

- Shaping policies: Data from legal and social research might inform policy decisions and ethical guidelines, ensuring that new technologies align with societal values and legal requirements.
- Addressing privacy concerns: Research findings might highlight key privacy issues and offer recommendations to protect individuals' rights while enabling responsible use of monitoring technologies.

Advancing research and innovation

- Driving interdisciplinary research: Data might facilitate collaboration among researchers from diverse fields such as healthcare, law, engineering, and sociology, fostering innovation in Active and Assisted Living (AAL) technologies.
- Enabling technology development: Technological data, such as sensor readings, audio recordings, and video datasets, might support the development and validation of

advanced algorithms for behaviour analysis, activity recognition, and privacy preservation.

- Benchmarking and validation: Existing and newly generated datasets might provide benchmarks for evaluating the performance and effectiveness of AAL technologies, ensuring they meet user needs and ethical standards.

Informing caregiver and stakeholder training

- Training programmes: Data on user experiences and technology effectiveness might contribute to training programmes for caregivers, healthcare professionals, and other stakeholders, helping them integrate AAL technologies into their work.
- Designing tailored solutions: Insights into specific challenges faced by caregivers and older adults might guide the customisation of technologies to better address individual needs.

Enhancing market reach

- Fostering user acceptance: Data on user perceptions and preferences might help identify and overcome barriers to acceptance, ensuring that technologies are designed with end-users in mind.
- Guiding commercialisation: Research findings might assist in developing commercially viable products by addressing ethical concerns, ensuring legal compliance, and improving usability.

1.8. Data aggregation

Data aggregation within GoodBrother projects might involve compiling, integrating, and processing data from diverse sources to support interdisciplinary research and innovation. By ensuring ethical compliance, transparency, and data harmonisation, aggregated datasets could provide a comprehensive foundation for analysis, serving as a valuable resource for understanding user needs, improving Active and Assisted Living (AAL) technologies, and fostering advancements in privacy-aware monitoring systems. The process would prioritise ethical considerations, data protection, and alignment with the FAIR (Findable, Accessible, Interoperable, Reusable) principles. The following outlines the purpose, methods, and potential outputs of data aggregation in GoodBrother projects.

Purpose of aggregation

Data aggregation might aim to:

- Combine diverse data sources: Integrate data from different disciplines (e.g., legal studies, social sciences, and technological projects) to provide holistic insights into user needs, privacy concerns, and system performance.
- Enable cross-domain analysis: Facilitate interdisciplinary research by creating datasets that can be used to address questions requiring inputs from multiple fields, such as the balance between user privacy and system functionality.

- Support model training and validation: In technological projects, aggregated datasets might be critical for training machine learning models and validating privacy-aware algorithms.

Methods of aggregation

Data aggregation might involve several techniques, depending on the research focus:

- Metadata mapping: Aggregating structured metadata (e.g., timestamps, activity labels) from different datasets to provide consistent annotations across projects.
- Cross-referencing data: Linking legal case studies, user feedback, and system performance metrics to examine the intersection of ethical, social, and technological factors.
- Data cleaning and standardisation: Ensuring that datasets collected from different sources are harmonised into compatible formats, such as .csv or .json, to facilitate analysis and sharing.
- Curated repositories: Developing curated repositories that catalogue datasets and software relevant to AAL tasks, including activity recognition, behaviour analysis, and privacy preservation.

Ethical and legal considerations

Data aggregation processes would comply with legal and ethical guidelines, ensuring that:

- GDPR compliance: Personal data is aggregated in a manner that respects the principles of data minimisation, anonymisation, and lawful processing.
- Transparency: Clear documentation would accompany aggregated datasets, describing their sources, methods of integration, and intended uses.
- Mitigation of bias: Steps would be taken to minimise biases that might arise from combining datasets with varying contexts, demographics, or objectives.

Outputs and accessibility

Aggregated data might contribute to:

- Curated data repositories: GoodBrother might provide a publicly accessible list of datasets and software tools relevant to audio- and video-based monitoring for AAL. Each dataset might be tagged by categories such as acquisition platform (e.g., audio, video, smart sensors), action focus (e.g., activity, behaviour, emotional state), and type of action (e.g., detection, monitoring, prevention).
- Research outputs: Aggregated data might enable the development of comprehensive analyses, fostering academic publications and collaborative projects across disciplines.

2. FAIR data

The GoodBrother COST Action aims to align with the FAIR (Findable, Accessible, Interoperable, Reusable) principles to ensure that the data generated or collected might be useful for current and future research while adhering to ethical and legal standards. By creating well-documented, easily accessible, and interoperable datasets, GoodBrother seeks to maximise the usability and impact of its research outputs. However, privacy concerns, particularly regarding sensitive audio and video data, might limit the availability of some datasets to external parties.

Datasets obtained from social science and health-related projects are generally not intended to be accessible to others outside the research team due to privacy concerns. Consequently, such data might not be deposited in public repositories unless specifically required by the project's goals. For these datasets, there might not be a need to ensure findability. Conversely, some datasets of images and videos collected in technological research projects might be made available to the research community. Licences for these datasets would be determined prior to publication to ensure responsible use.

GoodBrother has established a community on Zenodo (https://zenodo.org/communities/goodbrother) to store and share project-generated data. This platform might host image and video datasets, as well as reports, conference proceedings, and journal papers, promoting data sharing and dissemination where appropriate. The following outlines how GoodBrother might implement the FAIR principles[1], as stated at https://about.zenodo.org/principles:

- To be Findable:
    - F1: (meta)data are assigned a globally unique and persistent identifier
        - A DOI is issued to every published record on Zenodo.
    - F2: data are described with rich metadata (defined by R1 below)
        - Zenodo's metadata is compliant with DataCite's Metadata Schema[2] minimum and recommended terms, with a few additional enrichements.
    - F3: metadata clearly and explicitly include the identifier of the data it describes
        - The DOI is a top-level and a mandatory field in the metadata of each record.
    - F4: (meta)data are registered or indexed in a searchable resource

[1] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9.

[2] https://schema.datacite.org (last access: 07/11/2022)

        - Metadata of each record is indexed and searchable directly in Zenodo's search engine immediately after publishing.
        - Metadata of each record is sent to DataCite servers during DOI registration and indexed there.

- To be Accessible:
    - A1: (meta)data are retrievable by their identifier using a standardized communications protocol
        - Metadata for individual records as well as record collections are harvestable using the OAI-PMH[3] protocol by the record identifier and the collection name.
        - Metadata is also retrievable through the public REST API[4].
    - A1.1: the protocol is open, free, and universally implementable
        - OAI-PMH and REST are open, free, and universal protocols for information retrieval on the web.
    - A1.2: the protocol allows for an authentication and authorization procedure, where necessary
        - Metadata are publicly accessible and licensed under the public domain. No authorization is ever necessary to retrieve it.
    - A2: metadata are accessible, even when the data are no longer available
        - Data and metadata will be retained for the lifetime of the repository. This is currently the lifetime of the host laboratory CERN, which currently has an experimental program defined for the next 20 years at least.
        - Metadata are stored in high-availability database servers at CERN, which are separate from the data itself.
- To be Interoperable:
    - I1: (meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation.

[3] https://zenodo.org/oai2d (last access: 07/11/2022)
[4] https://developers.zenodo.org (last access: 07/11/2022)

        - Zenodo uses JSON Schema[5] as the internal representation of metadata and offers export to other popular formats such as Dublin Core[6] or MARCXML[7].
    - I2: (meta)data use vocabularies that follow FAIR principles
        - For certain terms, Zenodo refers to open, external vocabularies, e.g.: license (Open Definition[8]), funders (FundRef[9]) and grants (OpenAIRE[10]).
    - I3: (meta)data include qualified references to other (meta)data
        - Each referenced external piece of metadata is qualified by a resolvable URL.
- To be Reusable:
    - R1: (meta)data are richly described with a plurality of accurate and relevant attributes
        - Each record contains a minimum of DataCite's mandatory terms, with optionally additional DataCite recommended terms and Zenodo's enrichments.
    - R1.1: (meta)data are released with a clear and accessible data usage license
        - License is one of the mandatory terms in Zenodo's metadata and refers to an Open Definition[10] license.
        - Data downloaded by the users is subject to the license specified in the metadata by the uploader.
    - R1.2: (meta)data are associated with detailed provenance
        - All data and metadata uploaded is traceable to a registered Zenodo user.
        - Metadata can optionally describe the original authors of the published work.
    - R1.3: (meta)data meet domain-relevant community standards
        - Zenodo is not a domain-specific repository, yet through compliance with DataCite's Metadata Schema, metadata meets one of the broadest cross-domain standards available.

[5] https://json-schema.org (last access: 07/11/2022)
[6] http://dublincore.org (last access: 07/11/2022)
[7] https://www.loc.gov/marc/marcxml.html (last access: 07/11/2022)
[8] https://opendefinition.org (last access: 07/11/2022)
[9] https://www.crossref.org/services/funder-registry (last access: 07/11/2022)
[10] https://api.openaire.eu (last access: 07/11/2022)

Zenodo imposes certain limitations on the size of data that can be stored, with the current maximum size set at 50 GB per dataset. However, larger datasets might be accommodated upon request via Zenodo's contact form. If storing an extremely large dataset on Zenodo proves infeasible, alternative storage solutions would be explored. In such cases, the FAIR principles would remain a priority to ensure the data's usability, accessibility, and reusability.

For related projects that collect data, project funders or administrators might choose or be required to use other platforms or systems for data storage. These platforms might have specific requirements for licensing or distribution that take precedence over standard practices. In such instances, GoodBrother would reference these external projects and data sources rather than hosting or managing the data directly.

3. Allocation of resources

3.1. Financial costs

The financial costs associated with data management in GoodBrother projects might vary depending on the size and complexity of the datasets, as well as the specific requirements for implementing FAIR principles. In most cases, there are no direct costs for making data FAIR, especially when using platforms like Zenodo, which offers free storage for datasets under 50 GB. However, certain scenarios might incur additional costs:

- Storage for large datasets: Extremely large datasets, particularly those generated in technological projects, might exceed Zenodo's 50 GB limit. In such cases, alternative storage solutions, such as cloud-based repositories or institutional data centres, might require financial investment for additional capacity or specialised infrastructure.
- Data processing and preparation: Costs might arise for activities such as anonymisation, pseudonymisation, or standardisation of datasets to meet ethical and legal requirements. These activities could involve hiring data specialists or acquiring specialised software.
- Metadata generation: Developing detailed metadata to ensure compliance with FAIR principles might require dedicated time and resources, particularly for complex or multidisciplinary datasets.
- Licensing and access management: If datasets require specific licences or controlled access mechanisms (e.g., GDPR-compliant portals), associated legal or technical services might incur costs.
- Long-term preservation: Ensuring data availability over an extended period might involve fees for long-term repository use, especially if commercial or specialised platforms are chosen.

GoodBrother projects might rely on existing resources, such as institutional repositories and in-kind contributions from participating organisations, to minimise costs. For additional expenses, project budgets might include allocations for data management activities. In cases where external funding is required, GoodBrother might seek support from stakeholders, research grants, or other funding bodies.

3.2. Technical resources

The technical resources required for data management in GoodBrother projects might depend on the size, complexity, and type of data generated, as well as the specific requirements for adhering to FAIR principles. By leveraging robust storage solutions, processing tools, and metadata management systems, the projects aim to ensure compliance with FAIR principles while maintaining the integrity, security, and accessibility of the data. These resources might combine existing institutional and public infrastructure with specialised tools, as outlined below.

Most datasets might be stored using existing institutional or public repositories, such as Zenodo, which supports FAIR compliance for datasets up to 50 GB. For datasets exceeding Zenodo's capacity, additional storage solutions might include cloud-based services, such as AWS, Google Cloud, or Microsoft Azure, or high-capacity institutional servers. To ensure data integrity and recovery in case of hardware or system failures, redundant storage systems might be implemented. These measures would provide a scalable and reliable foundation for data management.

To process and standardise datasets into compatible formats such as .csv, .json, .mp4, or .wav, tools like R, Python, SPSS, and Excel might be employed. Ensuring compliance with GDPR and other privacy regulations might require software solutions for anonymisation or pseudonymisation. Open-source tools or specialised software could be used to securely prepare datasets for analysis and sharing while addressing privacy concerns.

For publicly accessible data, repositories like Zenodo might be the primary platform for storage and dissemination. However, for datasets requiring restricted access, secure portals with robust authentication mechanisms might be implemented to control data distribution. Additionally, open APIs might enable seamless data access and retrieval by the research community while maintaining compliance with licensing and privacy requirements.

Long-term data preservation might require specialised archival systems that ensure compliance with FAIR principles and maintain data usability even after the completion of active research projects. Solutions might also be selected to accommodate the gradual growth of datasets over time, particularly for ongoing or longitudinal studies. These systems would provide the scalability and durability necessary to manage large and complex datasets effectively.

4. Data security

Ensuring the security of data collected or generated within GoodBrother projects is a top priority, particularly given the sensitivity of some datasets, such as audio and video recordings, personal information, and health-related data. Robust measures would be implemented to safeguard data against unauthorised access, breaches, and misuse while complying with relevant legal and ethical standards, including the General Data Protection Regulation (GDPR). By combining technical measures like encryption and access control with secure infrastructure, proactive risk management, and regular audits, GoodBrother aims to ensure the

confidentiality, integrity, and availability of its datasets. The following outlines potential approaches to data security:

Security measures

To protect the confidentiality, integrity, and availability of data, the following technical and organisational measures might be employed:

- Encryption: All sensitive data, including personal information and multimedia files, might be encrypted during storage and transmission using state-of-the-art encryption protocols, such as AES-256.
- Access control: Access to datasets might be restricted to authorised personnel only, using role-based access controls (RBAC) and multi-factor authentication (MFA) to ensure that only approved individuals can access specific data.
- Anonymisation and pseudonymisation: Where feasible, datasets might be anonymised or pseudonymised to minimise the risk of identifying individuals. This process would ensure that personal data is protected even if access is inadvertently gained.
- Regular backups: Data backups might be performed regularly and stored securely in separate physical or virtual locations to protect against accidental loss or corruption.

Infrastructure and hosting

The choice of storage infrastructure would play a key role in ensuring data security:

- Secure repositories: Data might be stored in secure, GDPR-compliant repositories such as institutional servers, Zenodo, or trusted cloud providers like AWS, Microsoft Azure, or Google Cloud, all of which offer robust security features.
- Physical security: For data stored on-premises, physical security measures might include secure data centres with controlled access, surveillance, and environmental protections (e.g., fire suppression and temperature control).
- Cloud services: If cloud storage is used, providers would need to comply with international security standards, such as ISO 27001, to ensure the safety and integrity of stored data.

Compliance with GDPR

To meet GDPR requirements, data security measures would include:

- Data minimisation: Collecting and retaining only the data necessary for the specific purposes of the project, reducing the exposure of sensitive information.
- Data processing agreements: Ensuring that any external partners or service providers handling data sign data processing agreements (DPAs) to formalise their responsibilities under GDPR.
- Documentation: Maintaining detailed records of data security policies, practices, and any incidents to demonstrate accountability.

Risk management

A proactive approach to risk management might involve:

- Regular security audits: Periodic reviews of data storage systems and security practices might be conducted to identify vulnerabilities and ensure compliance with evolving standards.
- Incident response plans: Comprehensive incident response plans might be established to address potential breaches, including steps to mitigate harm, notify affected parties, and comply with reporting requirements.
- Training and awareness: Team members involved in data handling might receive regular training on best practices for data security and GDPR compliance.