Language Resources and Evaluation manuscript No.
(will be inserted by the editor)

Managing, storing, and sharing long-form recordings and their annotations

Lucas Gautheron · Nicolas Rochat · Alejandrina Cristia

Abstract The technique of long-form recordings via wearables is gaining momentum in different fields of research, notably linguistics and pathology. This technique, however, poses several technical challenges, some of which are amplified by the peculiarities of the data, including their sensitivity and their volume. In this paper, we begin by outlining key problems related to the management, storage, and sharing of the corpora that emerge when using this technique. We continue by proposing a multi-component solution to these problems, specifically in the case of daylong recordings of children. As part of this solution, we release *ChildProject*, a python package for performing the operations typically required by such datasets and for evaluating the reliability of annotations using a number of measures commonly used in speech processing and linguistics. Our proposal could be generalized to other populations.

Keywords daylong recordings, speech data management, data distribution, annotation evaluation, inter-rater reliability, reproducible research

1 Introduction

Long-form recordings are those collected over extended periods of time, typically via a wearable. Although the technique was used with normotypical adults decades ago (Mehl et al., 2001; Mehl and Pennebaker, 2003), it became widespread in the study of early childhood over the last decade since the publication of a seminal white paper by the LENA Foundation (Gilkerson and Richards, 2008). The LENA Foundation created a hardware-software combination that illuminated the potential of this technique for theoretical and applied purposes (e.g., Christakis et al. 2009; Warlaumont et al. 2014). More recently, long-form data is being discussed in the context of neurological disorders (e.g., Riad et al. 2020). In this article, we define the unique space of difficulties surrounding long-form recordings, and introduce a set of packages that provides practical solutions, with a focus on child-centered recordings. We end by discussing ways in which these solutions could be generalized to other populations. In order to demonstrate how our proposal could foster reproducible research on day-long recordings of children, we have released the source of the paper and the code used to build the figures which illustrate the capabilities of our python package in Section 4.

2 Problem space

Management of scientific data is a long-standing issue which has been the subject of substantial progress in the recent years. For instance, FAIR principles (Findability, Accessibility, Interoperability, and Reusability; see Wilkinson et al. 2016) have been proposed to help improve the usefulness of data and data analysis pipelines. Similarly, databases implementing these practices have emerged, such as Dataverse (King, 2007) and Zenodo (European Organization For Nuclear Research and OpenAIRE, 2013). The method of daylong recordings should incorporate such methodological advances. It should be noted, however, that some of the difficulties surrounding the management of corpora of daylong recordings are more idiosyncratic to this technique and therefore require specific solutions. Below, we list some of the challenges that researchers are likely to face while employing long-form recordings in naturalistic environments.

Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'Etudes cognitives, ENS, EHESS, CNRS, PSL University, Paris, France. E-mail: lucas.gautheron@gmail.com

	ACLEW starter	Van Dam
Audio's scope	5-minute clips	Full day
Automated annotations' format	none	LENA
Automated annotations' format	.eaf	.cha
Annotations' scope	only clips	Full day
Metadata	none	excel

Table 1: **Divergences between the Bergelson et al. (2017) and VanDam (2015) datasets**. Audios' scope indicates the size of the audio that has been archived: all recordings last for a full day, but for ACLEW starter, three five-minute clips were selected from each child. The automated annotations format indicates which software was used to annotate the audio automatically. Annotations' scope shows the scope of human annotation. Metadata indicates whether information about the children and recording were shared, and in what format.

The need for standards

Extant datasets rely on a wide variety of metadata structures, file formats, and naming conventions. For instance, some data from long-form recordings have been archived publicly on Databrary (such as the ACLEW starter set (Bergelson et al., 2017)) and HomeBank (including the VanDam Daylong corpus from VanDam 2015). Table 1 shows some divergence across the two, which is simply the result of researchers working in parallel. As a result of this divergence, however, each lab finds itself re-inventing the wheel. For instance, the HomeBankCode organization [1] contains at least 4 packages that do more or less the same operations, such as aggregating how much speech was produced in each recording, but implemented in different languages (MatLab, R, perl, and Python). This divergence may also hide different operationalizations, rendering comparisons across labs fraught, effectively diminishing replicability.[2]

Designing pipelines and analyses that are consistent across datasets requires standards for how the datasets are structured. Although this may represent an initial investment, such standards facilitate the pooling of research efforts, by allowing labs to benefit from code developed in other labs. Additionally, this field operates increasingly via collaborative cross-lab efforts. For instance, the ACLEW project[3] involved nine principal investigators (PIs) from five different countries, who needed a substantive initial investment to agree on a standard organization for the six corpora used in the project. We expect even larger collaborations to emerge in the future, a move that would benefit from standardization, as exemplified by the community that emerged around CHILDES for short-form recordings (MacWhinney, 2000a).

Keeping up with updates and contributions

Datasets are not frozen. Rather, they are continuously enriched through annotations provided by humans or new algorithms. Human annotations may also undergo corrections as errors are discovered. The process of collecting the recordings may also require a certain amount of time, as they are progressively returned by the field workers or the participants themselves. In the case of longitudinal studies, supplementary audio data may accumulate over several years. Researchers should be able to keep track of these changes while also upgrading their analyses. Moreover, several collaborators may be brought to contribute work to the same dataset simultaneously. To take the example of ACLEW, PIs first annotated a random selection of 2-minute clips for 10 children in-house. They then exchanged some of these audio clips so that the annotators in another lab could re-annotate the same data, for the purposes of inter-rater reliability. This revealed divergences in definitions, and all datasets needed to be revised. Finally, a second sample of 2-minute clips with high levels of speech activity were annotated, and another process of reliability was performed.

[1] `https://github.com/homebankcode/`

[2] *Replicability* is typically defined as the effort to re-do a study with a new sample, whereas *reproducibility* relates to re-doing the exact same analyses with the exact same data. Reproducibility is addressed in another section.

[3] `sites.google.com/site/aclewdid`

Delivering large amounts of data

Considering typical values for the bit depth and sampling rates of the recordings – 16 bits and 16 kilohertz respectively – yields a throughput of approximately three gigabytes per day of audio. Although there is a great deal of variation, past studies often involved at least 30 recording days (e.g., three days for each of ten children). The trend, however, is for datasets to be larger; for instance, last year, we collaborated in the collection of a single dataset, in which 200 children each contributed two recordings. Such datasets may exceed one terabyte. Moreover, these recordings can be associated with annotations spread across thousands of files. In the ACLEW example discussed above, there was one .eaf file per human annotator per type of annotation (i.e., random, high speech, random reliability, high speech reliability). In addition, the full day was analyzed with between one and four automated routines. Thus, for each recording day there were 8 annotation files, leading to 5 corpora $\times$ 10 children $\times$ 8 annotation = 400 annotation files. Other researchers will use one annotation file per clip selected for annotation, which quickly adds up to thousands of files. Even a small processing latency may result in significant overheads while gathering so many files.

Privacy

Long-form recordings are sensitive; they contain identifying and personal information about the participating family. In some cases, for instance if the family goes shopping and forgets to notify those around them, recordings could capture conversations which involve people who are unaware that they are being recorded. In addition, they may be subject to specific regulations, such as the European GDPR, the American HIPAA, and, depending on the place of collection and/or storage, laws on biometric data and incidental recording (which may vary across municipalities, states, and countries). For general ethical considerations, see Cychosz et al. (2020). Here, we focus on privacy in the context of complying with FAIR guidelines when using long-form recordings.

However, although long-form recordings are sensitive, many of the data types derived from them are not. With appropriate file-naming and meta-data practices, it is effectively possible to completely deidentify automated annotations (which at present never include automatic speech recognition). It is also often possible to deidentify human annotations, except when these involve transcribing what participants said, since participants will use personal names and reveal other personal details. Nonetheless, since this particular case involves a human doing the annotation, this human can be trained to modify the record (e.g., replace personal names with foils) and/or tag the annotation as sensitive and not to be openly shared. This is a practice called vetting, and it is one area in which the community working with long-form recordings has started to create standardized procedures, currently available from the HomeBank landing site (`homebank.talkbank.org`; e.g., VanDam et al. 2018).

Therefore, the ideal storing-and-sharing strategy should naturally enforce security and privacy safeguards by implementing access restrictions adapted to the level of confidentiality of the data. Data-access should be doable programmatically, and users should be able to download only the data that they need for their analysis.

Long-term availability

The collection of long-form recordings requires a considerable level of investment to explain the technique to families and communities, to ensure a secure data management system, and, in the case of remote populations, to access the site. In our experience, one data collection trip to a field site costs about 15 thousand US$.[4] These data are precious not only because of the investment that has gone into them, but also because they capture slices of life at a given point in time, which is particularly informative in the case of populations that are experiencing market integration or other forms of societal change – which today is most or all populations. Moreover, some communities who are collaborating in such research speak languages that are minority languages in the local context, and thus at a potential risk for being lost in the future. The conservation of naturalistic speech samples of children's language acquisition throughout a normal day could be precious for fueling future efforts of language revitalization (Nee, 2021). It would

[4] This grossly underestimates overall costs, because the best way to do any kind of field research is through maintaining strong bonds with the community and helping them in other ways throughout the year, not only during our visits (read more about ethical fieldwork on Broesch et al. 2020). A successful example for this is that of the UNM-UCSB Tsimane' Project (`http://tsimane.anth.ucsb.edu/`), which has been collaborating with the Tsimane' population since 2001. They are currently funded by a 5-year, 3-million US$ NIH grant `https://reporter.nih.gov/project-details/9538306`.

therefore be particularly damaging to lose such data prematurely, from financial, scientific, and human standpoints.

In addition, one advantage of daylong recordings over other observational methods such as parental reports is that they can be re-exploited at later times to observe behaviors that had not been foreseen at the time of data collection. This implies that their interest partly lies in long-term re-usability.

Moreover, even state-of-the-art speech processing tools still perform poorly on daylong recordings, due to their intrinsic noisy nature (Casillas et al., 2019). As a result, taking full advantage of present data will necessitate new or improved computational models, which may take years to develop. For example, the DIHARD Challenge series has been running for three consecutive years, and documents the difficulty of making headway with complex audio data (Ryant et al., 2018, 2019, 2020). For instance, the best submission for speaker diarization in their meeting subcorpus achieved about 35% Diarization Error Rate in 2018 and 2019, with improvements seen only in 2020, when the best system scored a 20% Diarization Error Rate (Neville Ryant, personal communication, 2021-04-09). Other tasks are progressing much more slowly. For instance, the best performance in a classifier for deciding whether adult speech was addressed to the child or to an adult scored about 70% correct in 2017 (Schuller et al., 2017) – but nobody has been able to beat this record since. Recordings should therefore remain available for long periods of time – potentially decades –, thus increasing the risk for data loss to occur at some point in their lifespan. For these reasons, the reliability of the storage design is critical, and redundancy is most certainly required. Likewise, persistent URLs may be needed in order to ensure the long-term accessibility of the datasets.

Findability

FAIR Principles include findability and accessibility. A crucial aspect of findability of datasets involves their being indexed in ways that potential re-users can discover them. Although we elaborate on it below, we want to already highlight HomeBank (`homebank.talkabank.org`) as one archiving option exists which is specific for long-form recordings, thus making any corpora hosted there easily discoverable by other researchers using the technique. Also of relevance is Databrary (`databrary.org`), an archive specialized on child development, which can thus make the data visible to the developmental science community. However, the current standard practice is archiving data in either one or another of these repositories, despite the fact that if a copy of the corpus were visible from one of these archives, the dataset would be overall more easily discovered. Additionally, it is uncertain whether these highly re-usable long-form recordings are visible to researchers who are more broadly interested in spoken corpora and/or naturalistic human behavior and/or other topics that could be studied in such data. In fact, one can conceive of a future in which the technique is used with people of different ages, in which case a system that allows users to discover other datasets based on relevant metadata would be ideal. For some research purposes (e.g., trying to stream overlapping voices and noise, technically referred to as "source separation") any recording may be useful, whereas for others (neurodegenerative disorders, early language acquisition) only some ages would. In any case, options exist to allow accessibility once a dataset is archived in one of those databases.

Reproducibility

Independent verification of results by a third party can be facilitated by improving the *reproducibility* of the analyses, i.e. by providing third-parties with enough data and information to re-derive claimed results. This itself maybe be challenging for a number of reasons, including the variety of software requirements, unclear data dependencies, or insufficiently documented steps. Sharing data sets and analyses is more complex than delivering a collection of static files; all the information that is necessary in order to re-execute any intermediate step of the analysis should also be adequately conveyed.

Current archiving options

The field of child-centered long-form recordings has benefited from a purpose-built scientific archive from an early stage. HomeBank VanDam et al. (2016) builds on the same architecture as CHILDES MacWhinney (2000b) and other TalkBank corpora. Although this architecture served the purposes of the language-oriented community well for short recordings, there are numerous issues when using it for long-form recordings. To begin with, curators do not directly control their datasets' contents and structures, and if a curator wants to make a modification, they need to ask the HomeBank management

team to make it for them. Similarly, other collaborators who spot errors cannot correct them directly, but again must request changes be made by the HomeBank management team. Only one type of annotation is innately managed, and that is CHAT MacWhinney (2000b), which is ideal for transcriptions of recordings. However, transcription is less central to studies of long-form audio.

As briefly noted above, Databrary `databrary.org` also already hosts some long-form recording data. The aforementioned ACLEW project actually committed to archiving data there, rather than on HomeBank, because it allowed direct control and update (without needing to ask the HomeBank management). As re-users, one of the most useful features of Databrary is the possibility to search the full archive for data pertaining to children of specific ages or origins. Using this archiving option led us to realize there were some limitations, including the fact that there is no API system, meaning that all updates need to be done via a graphical browser-based interface.

Additional options have been considered by researchers in the community, including OSF [5], and the Language Archive [6]. Detailing all their features is beyond the scope of the present paper, but some discussion can be found in Casillas et al. (2019).

Without denying their usefulness and importance, none of these archives provides perfect solutions to all of the problems we laid out above – and notably, in our vision, researchers should not have to choose among them when archiving their data. These limitations have brought us to envision a new strategy for sharing these datasets, which we detail next.

Our proposal

We propose a storing-and-sharing method designed to address the challenges outlined above simultaneously. It can be noted that these problems are, in many respects, similar to those faced by researchers in neuroimaging, a field which has long been confronting the need for reproducible analyses on large datasets of potentially sensitive data (Poldrack and Gorgolewski, 2014). Their experience may, therefore, provide precious insight for linguists, psychologists, and developmental scientists engaging with the big-data approach of long-form recordings. For instance, in the context of neuroimaging, Gorgolewski et al. (2016) have argued in favor of "machine-readable metadata", standard file structures and metadata, as well as consistency tests. Similarly, Eglen et al. (2017) have recommended the application of formatting standards, version control, and continuous testing.[7] In the following, we will demonstrate how all of these practices have been implemented in our proposed design. Albeit designed for child-centered daylong recordings, we believe our solution could be replicated across a wider range of datasets with constraints similar to those exposed above.

This solution relies on four main components, each of which is conceptually separable from the others: i) a standardized data format optimized for child-centered long-form recordings; ii) ChildProject, a python package to perform basic operations on these datasets; iii) DataLad, "a decentralized system for integrated discovery, management, and publication of digital objects of science" (Hanke et al., 2021a) iv) GIN, a live archiving option for storage and distribution. Our choice for each one of these components can be revisited based on the needs of a project and/or as other options appear. Table 2 summarizes which of these components helps address each of the challenges listed in Section 2.

3 Proposed solution

3.1 Dataset format

To begin with, we propose a set of proven standards which we use in the LAAC Team `https://lscp.dec.ens.fr/en/research/teams-lscp/language-acquisition-across-cultures` and which build on previous experience in several collaborative projects including ACLEW. It must be emphasized, however, that standards should be elaborated collaboratively by the community and that the following is merely a starting point.

[5] `osf.io`

[6] `https://archive.mpi.nl/tla/`

[7] Note that these concepts are all used in the key archiving options we evoked: HomeBank, Databrary, and the Language Archive all have defined metadata and file structures. However, they are *different* standards, which cannot be translated to each other, and which have not considered all the features that are relevant for long-form recordings, such as having multiple layers of annotations, with some based on sparse sampling. Additionally, the use of dataset versioning, automated consistency tests, and analyses based on subsumed datasets are less widespread in the language acquisition community.

Problem	**ChildProject** (Section 3.2)	**DataLad** (Section 3.3)	**GIN** (Section 3.4)
The need for standards	documented standards; tests; conversion routines		
Keeping up with updates and contributions		version control (git)	git repository host
Delivering large amounts of data	parallelised processing	git-annex	git-annex compatible; high storage capacity; parallelised operations
Ensuring privacy		private sub-datasets; private remotes; path-based or metadata-based storage rules;	Access Control Lists; SSH authentication
Long-term storage	tests (ensure integrity; detect missing files)	git; git-annex (remote synchronization, file availability and integrity checks, safe file deletion)	DOI registration
Findability	rich and standardized metadata	metadata aggregation metadata search	DOI registration; DataCite support repository search
Reproducibility		run/rerun/container-run functions	

Table 2: **Contributions of each component of our proposed design in resolving the difficulties caused by daylong recordings** and laid out in Section 2. ChildProject is a python package designed to perform recurrent tasks on the datasets; DataLad is a python package for the management of large, version-controlled datasets; GIN is a hosting provider dedicated to scientific data.

Data that are part of the same collection effort are bundled together within one folder[8], preferably a DataLad dataset (see Section 3.3). Datasets are packaged according to the structure given in fig. 1. The `metadata` folder contains at least three dataframes in CSV format: (i) `children.csv` contains information about the participants, such as their age or the language(s) they speak. (ii) `recordings.csv` contains the metadata for each recording, such as when the recording started, which device was used, or its relative path in the dataset. (iii) `annotations.csv` contains information concerning the annotations provided in the dataset, how they were produced, or which range they cover, etc. The dataframes are standardized according to guidelines which set conventional names for the columns and the range of allowed values. The guidelines are enforced through tests which print all the errors and inconsistencies in a dataset implemented in the ChildProject package introduced below.

The `recordings` folder contains two subfolders: `raw`, which stores the recordings as delivered by the experimenters, and `converted`, which contains processed copies of the recordings. All the audio files in `recordings/raw` are indexed in the recordings dataframe. Thus, there is no need for naming conventions for the audio files themselves, and maintainers can decide how they want to organize them.

The `annotations` folder contains all sets of annotations. Each set itself consists of a folder containing two subfolders: i) `raw`, which stores the output of the annotation pipelines and ii) `converted`, which stores the annotations after being converted to a standardized CSV format and indexed into `metadata/annotations.csv`. A set of annotations can contain an unlimited amount of subsets, with any amount of recursions. For instance, a set of human-produced annotations could include one subset per annotator. Recursion facilitates the inheritance of access permissions, as explained in Section 3.3.

[8] We believe a reasonable unit of bundling is the collection effort, for instance a single field trip, a full bout of data collection for a cross-sectional sample, or a set of recordings done more or less at the same time in a longitudinal sample. Given the possibilities of versioning, some users may decide they want to keep all data from a longitudinal sample in the same dataset, adding to it progressively over months and years, to avoid having duplicate children.csv files. That said, given DataLad's system of subdatasets (see Section 3.3), one can always define different datasets, each of which contains the recordings collected in subsequent time periods.

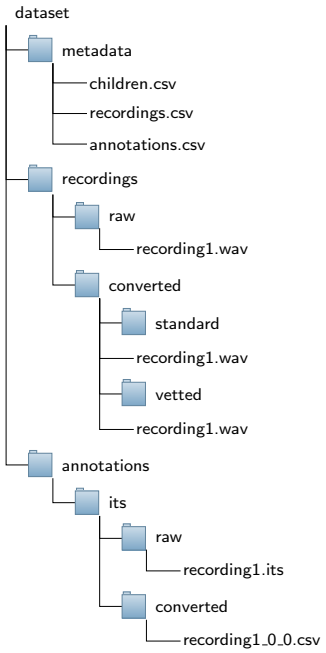


Fig. 1: **Structure of a dataset**. Metadata, recordings and annotations each belong to their own folder. Raw annotations (i.e., the audio files as they have been collected, before post-processing) are separated from their post-processed counterparts (in this case: standardized and vetted recordings). Similarly, raw annotations (in this case, LENA's its annotations) are set apart from the corresponding CSV version.

3.2 ChildProject

The ChildProject package is a Python 3.6+ package that performs common operations on a dataset of child-centered recordings. It can be used from the command-line or by importing the modules from within Python. Assuming the target datasets are packaged according to the standards summarized in Section 3.1, the package supports the functions listed below.

Listing errors and inconsistencies in a dataset

We provide a validation script that returns a detailed reporting of all the errors found within a dataset, such as violations of the formatting guidelines or missing files. Tests help enforce the standards that allow the commensurability of the datasets while guaranteeing the integrity and coherence of the data.

Converting and indexing annotations

The package converts input annotations to standardized, wide-table CSV dataframes. The columns in these wide-table formats have been determined based on previous work, and are largely specific to the goal of studying infants' language environment and production.

Annotations are indexed into a unique CSV dataframe which stores their location in the dataset, the set of annotations they belong to, and the recording and time interval they cover. The index, therefore, allows an easy retrieval of all the annotations that cover any given segment of audio, regardless of their original format and the naming conventions that were used. The system interfaces well with extant annotation standards. Currently, ChildProject supports: LENA annotations in .its (Xu et al., 2008); ELAN annotations following the ACLEW DAS template (Casillas et al. 2017, imported using Pympi: Lubbers and Torreira 2013-2021); CHAT annotations (MacWhinney, 2000b); as well as rttm files outputted by ACLEW tools, namely the Voice Type Classifier (VTC) by Lavechin et al. (2020), the Linguistic Unit Count Estimator (ALICE) by Räsänen et al. (2020), and the VoCalisation Maturity Network (VCMNet) by Futaisi et al. (2019). Users can also adapt routines for file types or conventions that vary. For instance, users can adapt the ELAN import developed for the ACLEW DAS template for their own template (e.g., `https://github.com/LAAC-LSCP/ChildProject/discussions/204`); and examples are also provided for Praat's .TextGrid files (Boersma, 2006). The package also supports custom, user-defined conversion routines.

Relying on the annotations index, the package can also calculate the intersection of the portions of audio covered by several annotators and align their annotations. This is useful when annotations from different annotators need to be combined (in order to retain the majority choice for instance) or compared (e.g., for reliability evaluations).

Choosing audio samples of the recordings to be annotated

As noted in the Introduction, recordings are too extensive to be manually annotated in their entirety. We and colleagues have typically annotated manually clips of 0.5-5 minutes in length, and the way these clips are extracted and annotated varies (as illustrated in Table 1).

The package allows the use of predefined or custom sampling algorithms. Samples' timestamps are exported to CSV dataframes. In order to keep track of the sample generating process, input parameters are simultaneously saved into a YAML file. Predefined samplers include a periodic sampler, a sampler targeting specific speakers' vocalizations, a sampler targeting regions of high-volubility according to input annotations, and a more agnostic sampler targeting high-energy regions. In all cases, the user can define the number of regions and their duration, as well as the context that may be inspected by human annotators. These options cover all documented sampling strategies.

Generating ELAN files ready to be annotated

Although there was some variability in terms of the program used for human annotation, the field has now by and large settled on ELAN (Wittenburg et al., 2006). ELAN employs xml files with a hierarchical structure which are both customizable and flexible. The ChildProject can be used to generate .eaf files which can be annotated with the ELAN software based on samples of the recordings drawn using the package, as described in Section 3.2.

Extracting and uploading audio samples to Zooniverse

The crowd-sourcing platform Zooniverse (Borne and Zooniverse Team, 2011) has been extensively employed in both natural (Zevin et al., 2017) and social sciences. More recently, researchers have been investigating its potential to classify samples of audio extracted from daylong recordings of children and the results have been encouraging (Semenzin et al., 2020a,b). We provide tools interfacing with Zooniverse's API for preparing and uploading audio samples to the platform and for retrieving the results, while protecting the privacy of the participants.

Audio processing

ChildProject allows the batch-conversion of the recordings to any target audio format (thanks to ffmpeg Developers 2021).

The package also implements a "vetting" (VanDam et al., 2018; Cychosz et al., 2020) pipeline, which mutes segments of the recordings previously annotated by humans as confidential while preserving the duration of the audio files. After being processed, the recordings can safely be shared with other researchers or annotators.

Another pipeline allows to perform filtering or linear combinations of audio channels for multi-channel recordings such as those produced with the BabyLogger[9]; if necessary, users can easily design custom audio converters suiting more specific needs.

Other functionalities

The package offers additional functions such as a pipeline that strips LENA's annotations from data that could be used to identify the participants, built upon previous code by MacEwan (2019).

Notably, the package facilitates the computation of a number of typical measures of annotations reliability and accuracy, as demonstrated in Section 4.

[9] `https://docs.babycloudlab.com/`

User empowerment

The present effort is led by one research team, and thus with personnel and funding that is not permanent. We therefore have done our best to provide information to help the community adopt and maintain this code in the future. Extensive documentation is provided on `https://childproject.readthedocs.io`, including detailed tutorials. The code is accessible on GitHub.com.

3.3 DataLad

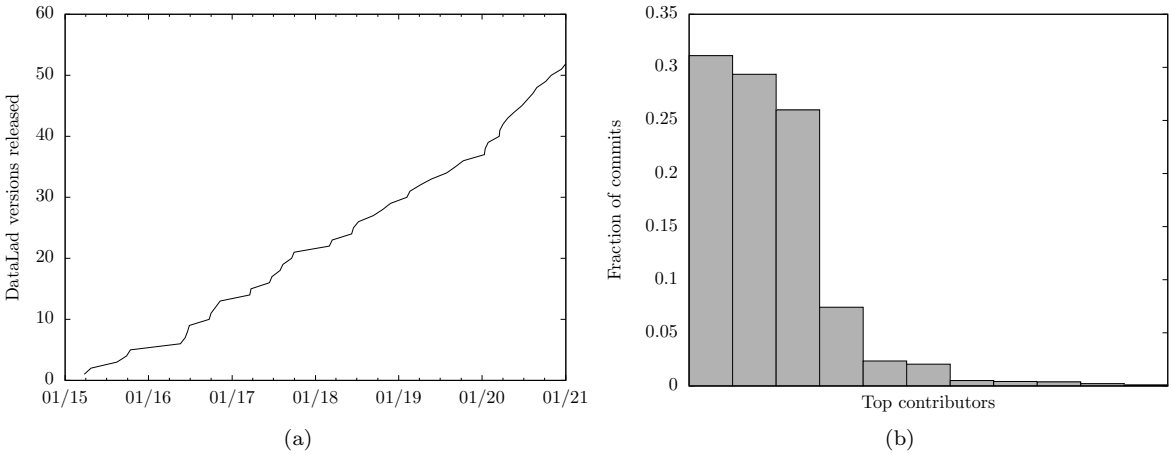


Fig. 2: **DataLad development activity**. (a) Amount of versions published across time. More than 50 versions have been released since 2015-01-01, at a steady pace. (b) Share of git commits held by top contributors in the last year (2020). At least three developers have contributed substantially, each of them being responsible for about 30% of the commits.

The combination of standards and the ChildProject package allows us to solve some of the problems laid out in the Introduction, but they do not directly provide solutions to the problems of data sharing and collaborative work. DataLad, however, has been specifically designed to address such needs.

DataLad (Wagner et al., 2020) was initially developed by researchers from the computational neuroscience community for the sharing of neuroimaging datasets. It has been under active development at a steady pace for at least six years (fig. 2a). It is co-funded by the United States NSF and the German Federal Ministry of Education and Research and has several major code developers (fig. 2b).

DataLad relies on git-annex, a software designed to manage large files with git. Over the years, git has rapidly overcome competitors such as Subversion, and it has been popularized by platforms such as GitLab and GitHub. However, git does not natively handle large binary files, our recordings included. Git-annex circumvents this issue by only versioning pointers to the large files. The actual content of the files is stored in an "annex". Annexes can be stored remotely on a variety of supports, including Amazon Glacier, Amazon S3, Backblaze B2, Box.com, Dropbox, FTP/SFTP, Google Cloud Storage, Google Drive, Internet Archive via S3, Microsoft Azure Blob Storage, Microsoft OneDrive, OpenDrive, OwnCloud, SkyDrive, Usenet, and Yandex Disk.

A DataLad dataset is, essentially, a git repository with an annex. As such, it naturally allows version control, easy collaboration with many contributors, and continuous testing. Furthermore, its use is intuitive to git users.

In using git-annex, DataLad enables users to download only the files that they need, without needing to fetch the whole dataset.

DataLad improves upon git-annex by adding a number of functionalities. One of them, dataset nesting, is built upon git submodules. A DataLad dataset can include sub-datasets, with as many levels of recursion as needed. This provides a natural solution to the question of how to document analyses, as an analysis repository can have the dataset on which it depends embedded as a subdataset. It also provides a good solution for the issue of different levels of data containing more or less identifying information, via the possibility of restricting permissions to different levels of the hierarchy.

Like git, DataLad is a decentralized system, meaning that data can be stored and replicated across several "remotes". DataLad authors have argued in favor of decentralized research data management, as it simplifies infrastructure migrations, and helps improve the scalibility of the data storage and distribution design Hanke et al. (2021b). Additionally, decentralization is notably useful in that it facilitates redundancy; files can be pushed simultaneously to several storage supports (e.g.: an external hard-drive, a cloud provider), thereby reducing the risk of data loss. In addition to that, when deleting large files from your local repository, DataLad will automatically make sure that more than a certain amount of remotes still own a copy the data, which by default is set to one.

Many of the *remotes* supported by DataLad require user-authentication, thus allowing for fine-grained access permissions management, such as Access-Control Lists (ACL). There are at least two ways to implement multiple levels of access within a dataset. One involves using sub-datasets with stricter access requirements. It is also possible to store data across several git-annex remotes with varying access permissions, depending on their sensitivity. Path-based pattern matching rules may configured in order to automatically select which remote the files should be pushed to. More flexible selection rules can be implemented using git-annex metadata, which allows to label files with `key=value` pairs. For instance, one could tag confidential files as `confidential=yes` and exclude these from certain remotes (blacklist). Alternatively, some files could be pushed to a certain remote provided they are labelled `public=yes` (whitelist).

DataLad's metadata[10] system can extract and aggregate information describing the contents of a collection of datasets. A search function then allows the discovery of datasets based on these metadata. We have developed a DataLad extension to extract meaningful metadata from datasets into DataLad's metadata system (Gautheron, 2021a). This allows, for instance, to search for datasets containing a given language. Moreover, DataLad's metadata can natively incorporate DataCite (Brase, 2010) descriptions into its own metadata.

DataLad may link data and software dependencies associated to a script as it is run. These scripts can later be re-executed by others, and the dependencies will automatically be downloaded. This way, DataLad can keep track of how each intermediate file was generated, thus simplifying the reproducibility of analyses. DataLad's handbook provides a tutorial to create a fully reproducible paper (Wagner et al., 2020, Chapter 22), and a template is available on GitHub (Wagner, 2020). The present paper has been built upon this template, and its source is available on GIN[11].

DataLad is domain-agnostic, which makes it suitable for maturing techniques such as language acquisition studies based on long-form recordings. The open-access data of the WU-Minn Human Connectome Project (Van Essen et al., 2013), totalling 80 terabytes to date, have been made available through DataLad [12].

3.4 Storage and distribution

DataLad does not provide, by itself, the infrastructure to share data. However, it allows maintainers to publish their content to a wide range of platforms. One can therefore implement different strategies for the storage and distribution of the data using any combination of these providers, depending on the constraints.

Table 3 sums up the most relevant characteristics of a few providers that are appropriate for our research, although many more could be considered. Datasets can only be cloned from providers that support git, and the large files can only be stored on those that support git-annex. Platforms that only support the former, such as GitHub, should therefore be used in tandem with providers that support the latter, like Amazon S3.

Among criteria of special interest are: the provider's ability to handle complex permissions; how much data it can accept; its ability to assign permanent URLs and identifiers to the datasets; and of course, whether it complies with the legislation regarding privacy. For our purposes, Table 3 suggests GIN is the best option, handling well large files, with highly customizable permissions, and Git-based version control and access (see Appendix A.3 for a practical use-case of GIN). That said, private projects are limited in space, although at the time of writing this limit can be raised by contributions to the GIN administrators. The next best option may be S3, and some users may prefer S3 when they do not have access to a local cluster, since S3 allows both easy storage and processing.

[10] `http://docs.datalad.org/en/stable/metadata.html`

[11] `https://gin.g-node.org/LAAC-LSCP/managing-storing-sharing-paper`

[12] `https://github.com/datalad-datasets/human-connectome-project-openaccess`

To render comparison of options easier, detailed examples of storage designs taken from real datasets are listed in Appendix A. Scripts to implement these strategies can be found on our GitHub and OSF (Gautheron, 2021b). We also provide a tutorial based on a public corpus (VanDam, 2015) to convert existing data to our standards and then publish it with DataLad[13]. We would like to emphasize that the flexibility of DataLad makes it very easy to migrate from one architecture to another. The underlying infrastructure may change, with little to no impact on the users, and little efforts from the maintainers.

In any case, we strongly recommend users to bear in mind that redundancy is important to make sure data are not lost, so a backup sibling may be hosted in an additional site (e.g., in a computer on campus in addition to the cloud-based version).

For instance, Perkel (2019) suggests several practices regarding backups, including automated backups, privacy safe-guarding, regular tests, and offline backups. Table 4 may orient the reader towards the functionalities of DataLad (and git-annex) which can be used to achieve these goals.

Provider	**Git**[a]	**Large files**[b]	**Authentication**	**Permissions**	**Quota**	**DOI registration**
SSH server	Yes	Yes	SSH	Unix	Self-hosted	No
GIN	Yes	Yes	HTTPS or SSH	ACL	[c]	Yes[c]
GitHub	Yes	No	HTTPS or SSH	ACL	–	No
GitLab	Yes	No	HTTPS or SSH	ACL	–	No
Nextcloud	No	Yes		ACL	Self-hosted	No
Amazon S3	No	Yes	API key+secret	IAM	Unlimited	No
OSF	Yes[d]	Yes[d]	Token	ACL	[e]	Yes

[a] The provider can store the git history and provide an URL from which the dataset can be installed.
[b] The provider handles git-annex large files.
[c] Contact the administrators
[d] With limitations (see `http://docs.datalad.org/projects/osf/en/latest/intro.html`)
[e] 5 GB for private projects, 50 GB for publics projects

Table 3: **Overview of several providers that can be used with DataLad**. The Unix permission system allows only one user and one group to be granted specific access rights. Access Control Lists (ACL) give more control, by enabling access to several groups and users. Amazon's Identity Access Management (IAM) can imitate ACLs, while providing more functionalities (fully-programmable; time-limited permissions; etc.)

Practice	Relevant software	Functionality
offline backups	DataLad	create-sibling, push[a]; export-archive;
	git-annex	copy;export[b]
backup automation	DataLad	siblings "publish-depends"[c]
privacy safe-guarding	git-annex	encryption
regular tests	git-annex	`fsck`[d]

Table 4: **Examples of recommended practices for data backups, associated to the software that could be used for their implementation**.

[a] creates a local sibling to which the data can be pushed, e.g. an external hard-drive.
[b] exports human-readable snapshots of a dataset
[c] "publish-depends" specifies which other siblings should be pushed to everytime some other sibling is updated. Maintainers can thus make sure that pushing to the main repository will trigger a push to the backup sibling.
[d] integrity check

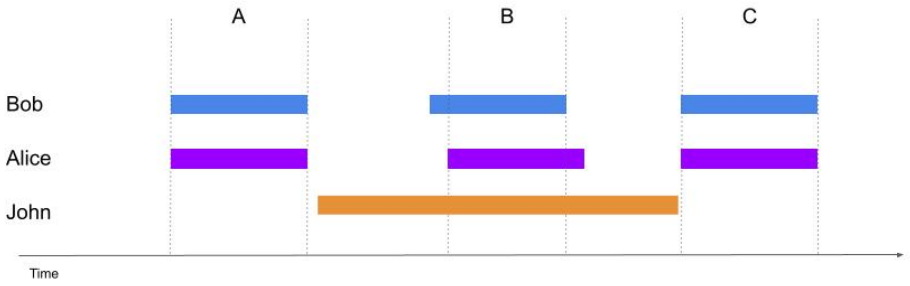
[13] `https://childproject.readthedocs.io/en/latest/vandam.html`

4 Application: evaluating annotations' reliability

Assessing the reliability of the annotations is crucial to linguistic research, but it can prove tedious in the case of daylong recordings. On one hand, analysis of the massive amounts of annotations generated by automatic tools may be computationally intensive. On the other hand, human annotations are usually sparse and thus more difficult to match with each other. Moreover, as emphasized in Section 2, the variety of file formats used to store the annotations makes it even harder to compare them.

Making use of the consistent data structures that it provides, the ChildProject package implements functions for extracting and aligning annotations regardless of their provenance or nature (human vs algorithm, ELAN vs Praat, etc.). It also provides functions to compute most of the metrics commonly used in linguistics and speech processing for comparing annotations, relying on existing efficient and field-tested implementations.

Figure 3 illustrates a recording annotated by three annotators (Alice, Bob and John). In this case, if one is interested in comparing the annotations by Bob and Alice, then the segments A, B and C should be compared. However, if the annotations common to all of the three annotators should be simultaneously compared, only the segment B should be considered. In real datasets with many recordings and several human and automatic annotators, the layout of annotations coverage may become much more complex. Relying on the index of annotations described in Section 3.2, the ChildProject package can calculate the intersection of the portions of audio covered by several annotators and return all matching annotations. These annotations can be filtered (e.g. excluding certain audio files), grouped according to certain characteristics (e.g. by participant), and aligned for subsequent analyses.



(a)

Fig. 3: **Example of time-intervals of a recording covered by three annotators**. Automated annotations usually cover whole recordings, while human annotators typically annotate periodic or targeted clips.

In psychometrics, the reliability of annotators is usually evaluated using inter-coder agreement indicators. The python package enables the calculation of some of these measures, including all of the coefficients implemented in the NLTK package (Loper and Bird, 2002) such as Krippendorff's Alpha (Krippendorff, 2013) and Fleiss' Kappa (Fleiss, 1971). The gamma method by Mathet et al. (2015), which aims to improve upon previous indicators by evaluating simultaneously the quality of both the segmentation and the categorization of speech, has been included *via* the `pygamma-agreement` package (Titeux and Riad, 2021).

It should be noted that these measures are most useful in the absence of ground truth, when reliability of the annotations can only be inferred by evaluating their overall agreement. Automatic annotators, however, are usually evaluated against a gold standard produced by human experts. In such cases, the package allows comparisons of pairs of annotators using metrics such as F-score, recall, and precision. Figure 4 illustrates this functionality. Additionally, the package can compute confusion matrices between two annotators, allowing more informative comparisons, as demonstrated in Figure 5. Finally, the python package interfaces well with `pyannote.metrics` (Bredin, 2017), and all the metrics implemented by the latter can be effectively used on the annotations managed with ChildProject.

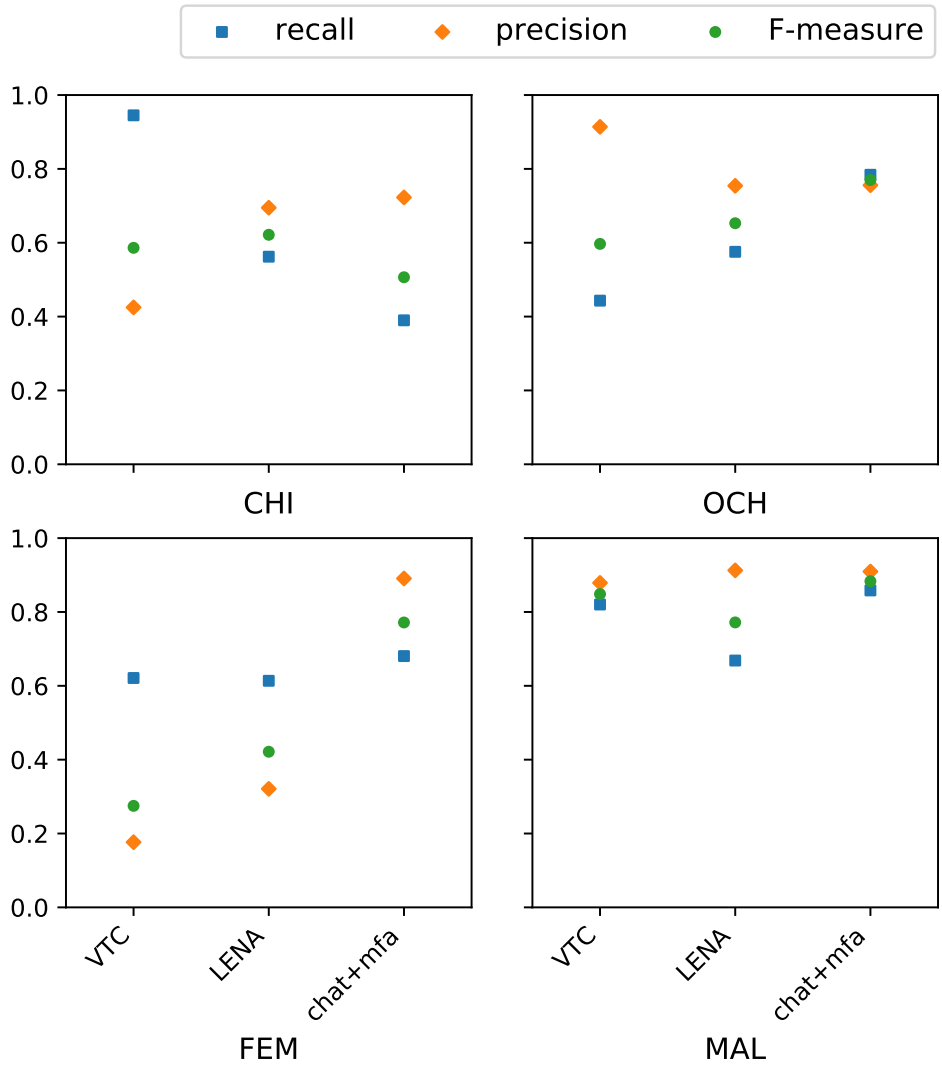


Fig. 4: **Examples of diarization performance evaluation using recall, precision and F1 score**. Audio from the the public VanDam corpus (VanDam, 2015) is annotated automatically according to who-speaks-when, using: the LENA diarizer; the Voice Type Classifier (VTC) by Lavechin et al. (2020); and manual CHAT transcriptions (MacWhinney, 2000b) adjusted with the Montreal Forced Aligner (McAuliffe et al., 2017) ("cha"). Speech segments are classified among four speaker types: the key child (CHI), other children (OCH), male adults (MAL) and female adults (FEM). Recall, precision and F1 score are calculated for each of these annotations, by comparing them to annotations of 5 × 1 minute clips annotated by a human annotator using ELAN ("eaf"; Wittenburg et al. 2006). The clips with the most adult words were targeted.

5 Generalization

The kinds of problems that our proposed approach addresses are relevant to at least three other bodies of data, all of them based on large datasets collected with wearables. First, there is a line of research on interaction and its effects on well-being among neurotypical adults (e.g., Mehl et al. (2001)). Second, audio data from wearables holds promise for individuals with medical and psychological conditions that have behavioral consequences which can evolve over time, including conditions that lead to coughing (Wu et al., 2018) and/or neurogenerative disorders (Riad et al., 2020). Third, some researchers hope to gather datasets on child development combining multiple information sources, such as parental reports, as well as other sensors picking up motion and psychophysiological data, with the goal of potentially intervening when it is needed (Levin et al., 2021).

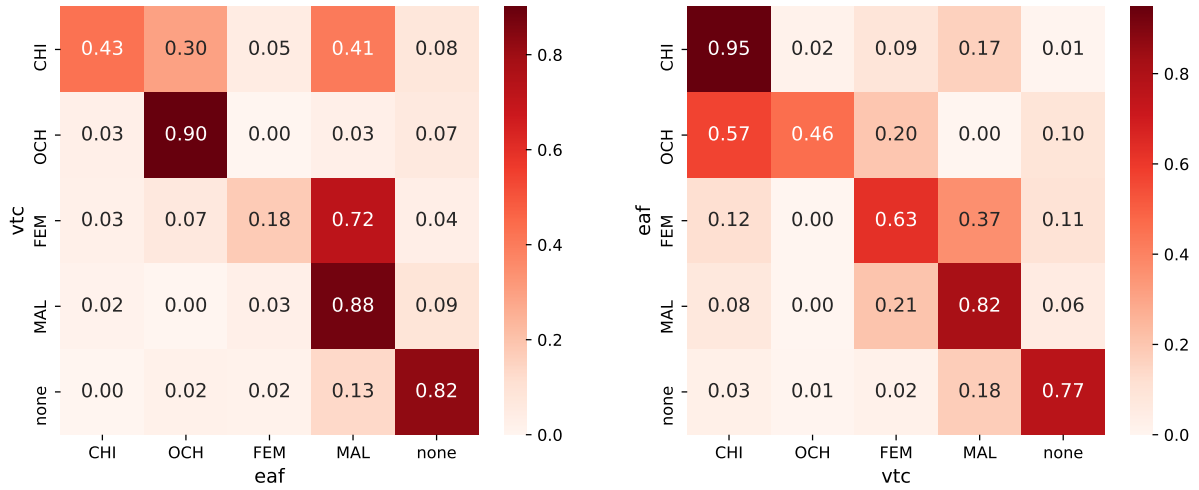


Fig. 5: **Example of diarization performance evaluation using confusion matrices** VTC annotations of the public VanDam corpus (VanDam, 2015) are compared to a gold standard manually annotated using ELAN (eaf). The first coefficient of the left side matrix should be read as: "43% of CHI segments from the VTC were also labelled as CHI by the human annotator" (i.e. as the precision). The first coefficient of the right side matrix should be read as: "95% of the portions labelled as CHI speech by the annotator were also labelled as CHI by the VTC" (i.e. as the recall). The sum of each row of the right-hand plot may exceed one due to overlapping speech. However, the diagonal should ideally be only ones.

Our proposed solution can be readily adapted to the first body of data: All that would need to be changed is renaming children.csv to participants.csv; renaming child_id to participant_id; and adapting which columns are mandatory and their format (e.g., it is cumbersome to express age in days for adults).

Generalizing our solution to the second body of data requires more adaptation. For such use cases, it would be ideal for the equipment to be left in the patients' house, so that it can be used for instance one day a week or month. Additional work is needed to facilitate this, ranging from making the equipment easier to use and more robust by for instance facilitating charging and secure data transfer from such off-site locations.

The third use case requires further adaptation, in addition to those just mentioned (making the sensors easy to use and allowing data transfer from potentially insecure home settings). In particular, multiple sensors' data need to be integrated together and time-stamped. We have made some progress in this sense in the context of the collection of multiple audio tracks collected with different physical devices (example forthcoming), but have not yet developed structure and code to support the integration of pictures, videos, heart rate data, parental questionnaire data, etc.

6 Limitations

DataLad and git-annex are well-documented, and, on the user's end, little knowledge beyond that of git is needed. Maintainers and resource administrators, however, will need a certain level of understanding in order to take full advantage of these tools. Recently, Powell (2021) has emphasized the shortcomings of decentralization and the inconveniences of a proliferation of databases with different access protocols. In the future, sharing data could be made even easier if off-the-shelf solutions compatible with DataLad were made readily available to linguists, psychologists, and developmental scientists. To this effect, we especially call for the attention of our colleagues working on linguistic databases. We are pleased to have found a solution on GIN – but it is possible that GIN administrators agreed to host our data because there is some potential connection with neuroimaging, whereas they may not be able to justify their use of resources for under-resourced languages and/or other projects that bear little connection to neuroimaging.

We should stress again that the use of the ChildProject package does not require the datasets to be managed with DataLad. They do need, however, to follow certain standards. Standards, of course, do not come without their own issues, especially in the present case of a maturing technique. They may

be challenged by ever-evolving software, hardware, and practices. However, we believe that the benefits of standardization outweigh its costs provided that it remains reasonably flexible. Such standards will further help to combine efforts from different teams across institutions. More procedures and scripts that solve recurrent tasks can be integrated into the ChildProject package, which might also speed up the development of future tools. One could argue that new proposed standards most usually end up increasing the amount of competing standards instead of bringing consensus. Nonetheless, if one were to eventually impose itself, well-structured datasets would still be easier to adapt than disordered data representations. Meanwhile, we look forward to discussing standards collaboratively with other teams via the GitHub platform, where anyone can create issues for improvements or bugs, submit pull-requests to integrate an improvement they have made, and/or have relevant conversations in the forum.

7 Summary

We provide a solution to the technical challenges related to the management, storage and sharing of datasets of child-centered daylong recordings. This solution relies on four components: i) a set of standards for the structuring of the datasets; ii) *ChildProject*, a python package to enforce these standards and perform useful operations on the datasets; iii) DataLad, a mature and actively developed version-control software for the management of scientific datasets; and iv) GIN, a storage provider compatible with Datalad. Building upon these standards, we have also provide tools to simplify the extraction of information from the annotations and the evaluation of their reliability along with the python package. The four components of our proposed design serve partially independent goals and can thus be decoupled, but we believe their combination would greatly benefit the technique of long-form recordings applied to language acquisition studies.

Declarations

Funding

This work has benefited from funding and/or institutional support from Agence Nationale de la Recherche (ANR-17-CE28-0007 LangAge, ANR-16-DATA-0004 ACLEW, ANR-14-CE30-0003 MechELex, ANR-17-EURE-0017); and the J. S. McDonnell Foundation Understanding Human Cognition Scholar Award. We also benefited from code developed in the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC), using the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. Additionally, we benefited from processing in GENCI-IDRIS, France (Grant-A0071011046). Some capabilities of our software depend on the Zooniverse.org platform, the development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation. The funders had no impact on this study.

Conflicts of interest/Competing interests

The authors have no conflict of interests to disclose.

Availability of data and material

This paper does not directly rely on specific data or material.

Code availability

The present paper can be reproduced from its source, which is hosted on GIN at `https://gin.g-node.org/LAAC-LSCP/managing-storing-sharing-paper`. The ChildProject package is available on GitHub at `https://github.com/LAAC-LSCP/ChildProject`. We provide scripts and templates for DataLad managed datasets at `http://doi.org/10.17605/OSF.IO/6VCXK` (Gautheron, 2021b). We also provide a DataLad extension to extract metadata from corpora of daylong recordings (Gautheron, 2021a).

A Examples of storage strategies

A.1 Example 1 - sharing a dataset within the lab

In the first example, Alice is hosting large datasets of a few terabytes of recordings and annotations and she wants to share them with Bob - a collaborator from her own institution - in a secure manner. Alice and Bob are familiar with GitHub, and they like its user-friendly features such as issues and pull requests. However, GitHub cannot handle such amounts of data.

Alice decides to store the git repository itself on GitHub – or a GitLab instance, it would not matter – thus allowing to benefit from their nice features while hosting the large files – the recordings and annotations – elsewhere. Alice's laboratory has its own cluster, with a large storage capacity. Thus, she decides to host the files there for free rather than using a Cloud provider.

Since Bob has been given SSH access to the cluster and belongs to the right UNIX group, he can download recordings and annotations from their joint institution cluster. Alice also made sure to configure the dataset in a way that makes sure every change published to GitHub is also published to the cluster, with DataLad's "publish-depends" option.

For backup purposes, a third sibling is hosted on Amazon S3 Glacier – which is cheaper than S3 at the expense of higher retrieval costs and delays – as a git-annex special remote. Special remotes do not store the git history and they cannot be used to clone the dataset. However, they can be used as a storage support for the recordings and other large files. In order to increase the security of the data, Alice uses encryption. Git-annex implements several encryption schemes[14]. The hybrid scheme allows to add public GPG keys at any time without additional decryption/encryption steps. Each user can then later decrypt the data with their own private key. This way, as long as at least one private GPG key has not been lost, data are still recoverable. This is especially valuable in that in naturally ensures redundancy of the decryption keys, which is critical in the case of encrypted backups.

By default, file names are hashed with an HMAC algorithm, and their content is encrypted with AES-128 – GPG's default –, although another algorithm could be selected.

This setup ensures redundancy of git files (hosted on both GitHub and the cluster) as well as large files (stored on both the cluster and Amazon Deep Glacier). It also allows Bob to signal and correct errors he finds, and/or to add annotations in a straightforward manner, benefiting Alice. By virtue of having siblings, they can make sure that their local dataset is organized in an identical manner, facilitating collaboration and reproducibility in their analyses.

Table 5 illustrates such a strategy. In this example, users install the dataset from a private GitHub repository. Continuous testing is configured with Travis CI[15], in order to ensure the integrity of the dataset at every step. GitHub Actions could also be used for that purpose[16].

We used this strategy – minus the Glacier backups – to maintain and deliver 4 datasets with 8700 hours of audio[17] for several months. The associated scripts can be found on Gautheron (2021b). We have now transitioned to using GIN for our main site, with our cluster as the backup. The scripts associated to this set-up can be found at the same location.

Sibling	Provider	Content	Access	Encryption
origin	GitHub	metadata; scripts	Lab	No
cluster	SSH server	everything	Lab	No
backup	Amazon Deep Glacier	recordings; annotations	Lab	AES-128

Table 5: Example 1 - Storage strategy example relying on GitHub and a cluster to deliver the data.

A.2 Example 2 - sharing large datasets with outside collaborators (S3)

The previous strategy is not suitable when complex permissions are required, since SSH remotes only handle Unix-style permissions (user, group, all).

Moreover, Alice may want to share the dataset with collaborators outside her lab, without giving them SSH access to its cluster. Or, she may not even own the infrastructure that would allow her to store and share such large amounts of data.

Instead, she decides to use Amazon S3 together with GitHub. Authorized users are provided their own Amazon S3 API key and secret, which are managed with Amazon's Identity and Access Manager (IAM). The GitHub is stripped from all confidential data, which are stored in the S3 annex only, allowing to manage access permissions entirely through IAM. This strategy is used by the Human Connectome Project[12].

Furthermore, Alice makes sure to encrypt GDPR relevant data, using strong symmetric encryption (AES-128). This strategy is illustrated in Table 6.

Amazon is superior to most alternatives for a number of reasons, including that it is highly tested, developed by engineers with a high-level of knowledge of the platform, and widely used. This means that the code is robust even before it is released, and it is widely tested once it is released. The fact that there are many users also entails that issues or questions can be looked up online. In addition, in the context of data durability, Amazon is a good choice because it is too big to fail, and thus probably available for the long-term. In addition, in sheer terms of flexibility and coverage, Amazon

[14] `https://git-annex.branchable.com/encryption/`

[15] https://travis-ci.com/

[16] https://docs.github.com/en/actions

[17] `https://github.com/LAAC-LSCP/datasets`

Sibling	Provider	Content	Access	Encryption
origin	GitHub	metadata; scripts	Collaborators	No
s3	Amazon S3	recordings; annotations	Collaborators	AES-128

Table 6: Example 2 - Storage strategy example relying on GitHub and Amazon S3.

provides a whole suite of tools (for data sharing, backups, and processing), which may be useful for researchers with little access to high-capacity infrastructures.

A.3 Example 3 - sharing large datasets with outside collaborators and multi-tier access (GIN)

Due to legislation in some countries, there are researchers who may not be authorized to store their data on Amazon. If they also do not have access to a local cluster (see Example 1) and/or even in the case that they have a local cluster, but need finer control of access permissions, there are alternatives which can be used as a workaround.

Finding herself in this setting, Alice decides to use the G-Node Infrastructure (GIN)[18], which is dedicated to providing "Modern Research Data Management for Neuroscience". GIN is similar to GitLab and GitHub in many aspects, except that is also supports git-annex and thus can directly host the large files that required third-party providers while using those platforms.

Just like GitLab or GitHub, it can handle complex permissions, at the user or group-level, thus surpassing Unix-style permissions management.

In this case, Alice needs three permission tiers: 1) read-only access to anonymized data, 2) read-only access to confidential data, and 3) read and write access to the whole data. In order to achieve this, she creates two GIN siblings per dataset: `origin` and `confidential`. The dataset is configured to publish all the files whose path contains `/confidential/` to the `confidential` repository, while the rest of the data is published to `origin`. Alice could then great read-only access to `origin` to both Bob and Carol, while restricting the access to `confidential` to Bob only.

Since Alice has not been allowed to use a cloud provider, and is lacking a local infrastructure, she needs an alternate solution for her backups. She may use external hard drives, as DataLad allows to push data to a local storage as with any other kind of storage.

Table 7 sums up this strategy, which is currently used to deliver the EL1000 dataset[19] – except for the backup, which is located at our cluster –. The EL1000 is a composite dataset, created by the contribution of 18 different teams that collected data independently but using comparable methods.

Sibling	Provider	Content	Access	Encryption
origin	GIN	files NOT matching `**/confidential/*`	Alice (read+write); Bob, Carol (read-only)	No
confidential	GIN	files matching `**/confidential/*`	Alice (read+write); Bob (read-only)	No
backup	external hard drive	everything	Alice	No

Table 7: Example 3 - Storage strategy example relying solely on GIN to deliver the data.

A.4 Example 4 - Sharing smaller datasets (OSF)

The Open Science Framework (OSF) is especially interesting because it supports DOI registration, providing permanent URLs to access the datasets. Moreover, an extension of DataLad has specifically been developed to work with OSF, which may host both the git repository and the large files (see Table 3). In addition, Shibboleth credentials can be used with OSF.

Low quotas are an important downside with OSF. Public projects are limited to 50 GB, and private projects cannot exceed 5 GB, which is too low for most long-form datasets. However, OSF could be used only to host the git repository, effectively providing a permanent URL from which the dataset can be installed, as long as the content of the large files remains available from a third-party provider, e.g. with Amazon S3. Table 8 illustrates such a strategy.

We use a reversed approach for our demo dataset[20] based on (VanDam, 2015), by hosting the git repository on GitHub, and hosting the large files on OSF. This is possible only because of the small size of the dataset.

[18] `https://gin.g-node.org/`

[19] `https://gin.g-node.org/EL1000/EL1000`

[20] https://github.com/LAAC-LSCP/vandam-daylong-demo

Sibling	Provider	Content	Access	Encryption
origin	OSF	metadata; scripts	Everyone	No
s3	Amazon S3	annotations; recordings	Alice, Bob and Carol	No

Table 8: Example 4 - Storage strategy example relying on OSF and Amazon S3 to deliver the data.

References

Bergelson E, Warlaumont A, Cristia A, Casillas M, Rosemberg C, Soderstrom M, Rowland C, Durrant S, Bunce J (2017) Starter-aclew. DOI 10.17910/B7.390, URL http://databrary.org/volume/390

Boersma P (2006) Praat: doing phonetics by computer. http://www praat org/

Borne KD, Zooniverse Team (2011) The Zooniverse: A Framework for Knowledge Discovery from Citizen Science Data. In: AGU Fall Meeting Abstracts, vol 2011, pp ED23C–0650

Brase J (2010) Datacite - a global registration agency for research data. SSRN Electronic Journal DOI 10.2139/ssrn.1639998, URL https://doi.org/10.2139/ssrn.1639998

Bredin H (2017) pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. In: Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, URL http://pyannote.github.io/pyannote-metrics

Broesch T, Crittenden AN, Beheim BA, Blackwell AD, Bunce JA, Colleran H, Hagel K, Kline M, McElreath R, Nelson RG, et al. (2020) Navigating cross-cultural research: methodological and ethical considerations. Proceedings of the Royal Society B 287(1935):20201245

Casillas M, Bergelson E, Warlaumont AS, Cristia A, Soderstrom M, VanDam M, Sloetjes H (2017) A new workflow for semi-automatized annotations: Tests with long-form naturalistic recordings of childrens language environments. In: Proc. Interspeech 2017, pp 2098–2102, DOI 10.21437/Interspeech.2017-1418, URL http://dx.doi.org/10.21437/Interspeech.2017-1418

Casillas M, Cristia A, Zwaan R, Dingemanse M (2019) A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings. Collabra: Psychology 5(1), DOI 10.1525/collabra.209, URL https://doi.org/10.1525/collabra.209, 24, https://online.ucpress.edu/collabra/article-pdf/5/1/24/437539/209-3199-1-pb.pdf

Christakis DA, Gilkerson J, Richards JA, Zimmerman FJ, Garrison MM, Xu D, Gray S, Yapanel U (2009) Audible television and decreased adult words, infant vocalizations, and conversational turns: a population-based study. Archives of pediatrics & adolescent medicine 163(6):554–558

Cychosz M, Romeo R, Soderstrom M, Scaff C, Ganek H, Cristia A, Casillas M, de Barbaro K, Bang JY, Weisleder A (2020) Longform recordings of everyday life: Ethics for best practices. Behavior Research Methods 52(5):1951–1969, DOI 10.3758/s13428-020-01365-9, URL https://doi.org/10.3758/s13428-020-01365-9

Eglen SJ, Marwick B, Halchenko YO, Hanke M, Sufi S, Gleeson P, Silver RA, Davison AP, Lanyon L, Abrams M, Wachtler T, Willshaw DJ, Pouzat C, Poline JB (2017) Toward standard practices for sharing computer code and programs in neuroscience. Nature Neuroscience 20(6):770–773, DOI 10.1038/nn.4550, URL https://doi.org/10.1038/nn.4550

European Organization For Nuclear Research, OpenAIRE (2013) Zenodo. DOI 10.25495/7GXK-RD71, URL https://www.zenodo.org/

ffmpeg Developers (2021) ffmpeg tool. URL http://ffmpeg.org/

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychological Bulletin 76(5):378–382, DOI 10.1037/h0031619, URL https://doi.org/10.1037/h0031619

Futaisi NA, Zhang Z, Cristia A, Warlaumont A, Schuller B (2019) VCMNet: Weakly supervised learning for automatic infant vocalisation maturity analysis. In: 2019 International Conference on Multimodal Interaction, ACM, DOI 10.1145/3340555.3353751, URL https://doi.org/10.1145/3340555.3353751

Gautheron L (2021a) Datalad extension for child-centered in-situ recordings DOI 10.17605/OSF.IO/C2J5A, URL https://osf.io/c2j5a/

Gautheron L (2021b) Datalad procedures for the management of long-form recordings DOI 10.17605/OSF.IO/6VCXK, URL https://osf.io/6vcxk/

Gilkerson J, Richards J (2008) The power of talk (lena foundation technical report ltr-01-2)

Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, Das S, Duff EP, Flandin G, Ghosh SS, Glatard T, Halchenko YO, Handwerker DA, Hanke M, Keator D, Li X, Michael Z, Maumet C, Nichols BN, Nichols TE, Pellman J, Poline JB, Rokem A, Schaefer G, Sochat V, Triplett W, Turner JA, Varoquaux G, Poldrack RA (2016) The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Scientific Data 3(1), DOI 10.1038/sdata.2016.44, URL https://doi.org/10.1038/sdata.2016.44

Hanke M, Pestilli F, Wagner AS, Markiewicz CJ, Poline JB, Halchenko YO (2021a) In defense of decentralized research data management. Neuroforum 0(0):000010151520200037, DOI 10.1515/nf-2020-0037, URL https://www.degruyter.com/document/doi/10.1515/nf-2020-0037/html

Hanke M, Pestilli F, Wagner AS, Markiewicz CJ, Poline JB, Halchenko YO (2021b) In defense of decentralized research data management. Neuroforum 0(0), DOI 10.1515/nf-2020-0037, URL https://doi.org/10.1515/nf-2020-0037

King G (2007) An introduction to the dataverse network as an infrastructure for data sharing. Sociological Methods and Research 36:173–199

Krippendorff K (2013) Content analysis : an introduction to its methodology. SAGE, Los Angeles London

Lavechin M, Bousbib R, Bredin H, Dupoux E, Cristia A (2020) An open-source voice type classifier for child-centered daylong recordings. Interspeech

Levin HI, Egger D, Andres L, Johnson M, Bearman SK, de Barbaro K (2021) Sensing everyday activity: Parent perceptions and feasibility. Infant Behavior and Development 62:101511

Loper E, Bird S (2002) Nltk: The natural language toolkit. CoRR cs.CL/0205028, URL http://dblp.uni-trier.de/db/journals/corr/corr0205.html#cs-CL-0205028

Lubbers M, Torreira F (2013-2021) pympi-ling: a Python module for processing ELANs EAF and Praats TextGrid annotation files. https://pypi.python.org/pypi/pympi-ling, version 1.70

MacEwan S (2019) Homebank its file anonymizer. URL https://github.com/HomeBankCode/ITS_annonymizer

MacWhinney B (2000a) The CHILDES project: The database, vol 2. Psychology Press

MacWhinney B (2000b) The CHILDES project: Tools for analyzing talk (third edition): Volume i: Transcription format and programs, volume II: The database. Computational Linguistics 26(4):657–657, DOI 10.1162/coli.2000.26.4.657, URL https://doi.org/10.1162/coli.2000.26.4.657

Mathet Y, Widlöcher A, Métivier JP (2015) The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. Computational Linguistics 41(3):437–479, DOI 10.1162/coli_a_00227, URL https://doi.org/10.1162/coli_a_00227

McAuliffe M, Socolof M, Mihuc S, Wagner M, Sonderegger M (2017) Montreal forced aligner: Trainable text-speech alignment using kaldi. In: Proc. Interspeech 2017, pp 498–502, DOI 10.21437/Interspeech.2017-1386, URL http://dx.doi.org/10.21437/Interspeech.2017-1386

Mehl MR, Pennebaker JW (2003) The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. Journal of Personality and Social Psychology 84(4):857–870, DOI 10.1037/0022-3514.84.4.857, URL https://doi.org/10.1037/0022-3514.84.4.857

Mehl MR, Pennebaker JW, Crow DM, Dabbs J, Price JH (2001) The electronically activated recorder (EAR): A device for sampling naturalistic daily activities and conversations. Behavior Research Methods, Instruments, & Computers 33(4):517–523, DOI 10.3758/bf03195410, URL https://doi.org/10.3758/bf03195410

Nee J (2021) Understanding the effects of language revitalization workshops using long-format speech environment recordings. Proceedings of the Linguistic Society of America 6(1):213, DOI 10.3765/plsa.v6i1.4967, URL https://doi.org/10.3765/plsa.v6i1.4967

Perkel JM (2019) 11 ways to avert a data-storage disaster. Nature 568(7750):131–132, DOI 10.1038/d41586-019-01040-w, URL https://doi.org/10.1038/d41586-019-01040-w

Poldrack RA, Gorgolewski KJ (2014) Making big data open: data sharing in neuroimaging. Nature Neuroscience 17(11):1510–1517, DOI 10.1038/nn.3818, URL https://doi.org/10.1038/nn.3818

Powell K (2021) The broken promise that undermines human genome research. Nature 590(7845):198–201, DOI 10.1038/d41586-021-00331-5, URL https://doi.org/10.1038/d41586-021-00331-5

Räsänen O, Seshadri S, Lavechin M, Cristia A, Casillas M (2020) Alice: An open-source tool for automatic measurement of phoneme, syllable, and word counts from child-centered daylong recordings. Behavior Research Methods pp 1–18

Riad R, Titeux H, Lemoine L, Montillot J, Bagnou JH, Cao XN, Dupoux E, Bachoud-Lévi AC (2020) Vocal markers from sustained phonation in huntington's disease. Interspeech

Ryant N, Church K, Cieri C, Cristia A, Du J, Ganapathy S, Liberman M (2018) First dihard challenge evaluation plan. 2018, tech Rep

Ryant N, Church K, Cieri C, Cristia A, Du J, Ganapathy S, Liberman M (2019) The second dihard diarization challenge: Dataset, task, and baselines. arXiv preprint arXiv:190607839

Ryant N, Church K, Cieri C, Du J, Ganapathy S, Liberman M (2020) Third dihard challenge evaluation plan. arXiv preprint arXiv:200605815

Schuller B, Steidl S, Batliner A, Bergelson E, Krajewski J, Janott C, Amatuni A, Casillas M, Seidl A, Soderstrom M, et al. (2017) The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring. In: Interspeech

Semenzin C, Hamrick L, Seidl A, Lynne Kelleher B, Cristia A (2020a) Describing vocalizations in young children: A big data approach through citizen science annotation DOI 10.31219/osf.io/z6exv, URL https://doi.org/10.31219/osf.io/z6exv

Semenzin C, Hamrick L, Seidl A, Lynne Kelleher B, Cristia A (2020b) Towards large-scale data annotation of audio from wearables: Validating zooniverse annotations of infant vocalization types DOI 10.31219/osf.io/gpxf5, URL https://doi.org/10.31219/osf.io/gpxf5

Titeux H, Riad R (2021) pygamma-agreement: Gamma γ measure for inter/intra-annotator agreement in Python, URL https://hal.archives-ouvertes.fr/hal-03144116, working paper or preprint

Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K, Consortium ftWMH (2013) The wu-minn human connectome project: An overview. NeuroImage 80:62–79, DOI 10.1016/j.neuroimage.2013.05.041, URL http://europepmc.org/articles/pmc3724347?pdf=render

VanDam M (2015) Homebank vandam public 5-minute corpus. DOI 10.21415/T5388S, URL http://homebank.talkbank.org/access/Public/VanDam-5minute.html

VanDam M, Warlaumont AS, Bergelson E, Cristia A, Soderstrom M, De Palma P, MacWhinney B (2016) Homebank: An online repository of daylong child-centered audio recordings. In: Seminars in Speech and Language, NIH Public Access, vol 37, p 128

VanDam M, Warlaumont A, MacWhinney B, Soderstrom M, Bergelson E (2018) Vetting manual: Preparation of recordings for unrestricted publication in homebank (version 1.1). DOI: https://doi org/1021415/T56H4M

Wagner A (2020) datalad-handbook/repro-paper-sketch: A template to create a reproducible paper with latex, makefiles, python, and datalad. https://github.com/datalad-handbook/repro-paper-sketch/, (Accessed on 04/30/2021)

Wagner AS, Waite LK, Meyer K, Heckner MK, Kadelka T, Reuter N, Waite AQ, Poldrack B, Markiewicz CJ, Halchenko YO, Vavra P, Chormai P, Poline JB, Paas LK, Herholz P, Mochalski LN, Kraljevic N, Wiersch L, Hutton A, Hanke M (2020) The DataLad Handbook. Zenodo, DOI 10.5281/ZENODO.3608612, URL https://zenodo.org/record/3608612

Warlaumont AS, Richards JA, Gilkerson J, Oller DK (2014) A social feedback loop for speech development and its reduction in autism. Psychological science 25(7):1314–1324

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3(1), DOI 10.1038/sdata.2016.18, URL https://doi.org/10.1038/sdata.2016.18

Wittenburg P, Brugman H, Russel A, Klassmann A, Sloetjes H (2006) Elan: a professional framework for multimodality research. In: 5th International Conference on Language Resources and Evaluation (LREC 2006), pp 1556–1559

Wu R, Liaqat D, de Lara E, Son T, Rudzicz F, Alshaer H, Abed-Esfahani P, Gershon AS (2018) Feasibility of using a smartwatch to intensively monitor patients with chronic obstructive pulmonary disease: Prospective cohort study. JMIR mHealth and uHealth 6(6):e10046, DOI 10.2196/10046, URL https://doi.org/10.2196/10046

Xu D, Yapanel U, Gray S, Baer C (2008) The lenatm language environment analysis system: The interpretive time segments (its) file. LENA Research Foundation Technical Report LTR-04-2

Zevin M, Coughlin S, Bahaadini S, Besler E, Rohani N, Allen S, Cabero M, Crowston K, Katsaggelos AK, Larson SL, Lee TK, Lintott C, Littenberg TB, Lundgren A, Østerlund C, Smith JR, Trouille L, Kalogera V (2017) Gravity spy: integrating advanced LIGO detector characterization, machine learning, and citizen science. Classical and Quantum Gravity 34(6):064003, DOI 10.1088/1361-6382/aa5cea, URL https://doi.org/10.1088/1361-6382/aa5cea