Notes on Corpus Construction

R.J.J.H. van Son

Contents

Notes on Corpus Construction

R.J.J.H. van Son
AVL, Amsterdam & ACLC/IFA, University of Amsterdam
the Netherlands
R.J.J.H.vanSon@gmail.com

30 October 2017

Abstract

Corpora of spoken language are both important for research, rare, and frequently difficult to use. These notes give some suggestions to ease the construction and distribution of corpora.

1 Introduction

During short term research projects there tends to come a moment where the question is raised whether and how to preserve the data gathered over the years. The present notes are intended to help researchers who want to preserve the data gathered in their projects and make these data available to other researchers. I want to suggest some tips that might help to improve the availability of original research data and ease the efforts needed to make them available. There are good books that go into detail about the planning and construction of language corpora [1, 2, 3] as well as conferences, e.g., LREC, and specialized workshops [4, 5, 6, 7]. Some general, common sense, rules for handling digital data can be found in [8]. The current notes will be limited to practical tips on how to convert existing, small, data sets into a corpus.

The focus of these notes will be on mostly static data from small and short term research projects, e.g., by PhD students or Post-Docs, but the suggestions can easily be adapted to other situations. In such projects, the amount of data is limited and the time available to organize them is even more limited. The data should be organized in such a way that it can be stored indefinitely on a single departmental server. However, it should be easy to add more data to an existing corpus or to copy and move the whole collection from one location to another. For an example of how to construct a much larger corpus, please see the *Spoken Dutch Corpus* [9, 10] or the *DOBES* project [11].

These notes will discuss corpus construction based on a few example spoken language corpora that are available from the IFA Spoken Language Corpora [12, 13, 14, 15]. The topics that will be discussed are: Corpus structure, content and documentation, and distribution. These topics will be illustrated with a few examples

2 Corpus Structure

We assume a corpus that consists of files which are stored in directories. A corpus is a storage structure for (primary) data. As such it should help in bookkeeping of experimental data. The first question will be where to store the resulting corpus. The best location would be on a well maintained web server or the "cloud" (e.g., Dropbox [16]). In one stroke, such on-line storage would take care of many chores, from back-ups to distribution. But if such a solution is not available, you can start small, just a hard drive to work on is enough, and preferably, a second drive to back up to.

2.1 Recordings and media

The first decision to make when constructing a corpus is *What data should be stored in the corpus?* We can make a distinction between original media, e.g., texts, audio and video recordings, but also EEG recordings, and data connected to these media, e.g., annotations, subject responses or evaluations, and analysis results.

Beside these data, there is also metadata about the subjects involved and experimental circumstances. These can most easily be stored in a separate section of the corpus.

A decision that has to be made on this topic is the *unit of data* that should be stored. As everything is stored as files, each file should contain a *unit of data*. This can be a text, or a text fragment. It can be a sound recording, but also sections of a sound recording or even individual utterances or words. Storing several copies or parts of each item complicates the corpus considerably. It becomes time consuming to ensure that all instances of a certain item are correctly selected and updated when there is a correction or replacement. And it is almost inevitable that there will be corrections and replacements at some stage in the building and use of a corpus.

There is one exception to the advice to keep only single instances of every data item. When a corpus is constructed by processing original materials, e.g., edited video recordings or large texts, then it is advisable to keep archival copies of the original materials. However, these archival copies do not have to be part of the *real* corpus. This is especially important for video and audio recordings. It is very likely that such recordings will not be used in the format they were recorded in. Very often, video or audio compression has to be applied. The format conversions can lead to artifacts or loss of information [17]. Accessing the original data can then be necessary to be able to decide how to interpret ambiguous or suspicious observations.

2.2 Directories

The corpus will be organized in directories. Files of a single type that belong together will be stored in a single directory. This means, that in a video dialog corpus, all video recordings can be stored in a single directory. The extracted audio files will be stored in a different directory, as will be the transliterations of the speech and the annotations of, e.g., gaze direction. In more complex corpora, it could be better to group files in sub-, or sub-sub-directories. So, a corpus of pathological speech build from a group of separate studies could have separate sub-directories for different pathologies, and sub-sub-directories of the different tasks which were recorded, and sub-sub-subdirectories for the different studies from which the recording were taken.

The easiest way to organize such a complex corpus would be to make mirror directory trees for the different types of recordings (media), and transliterations and annotations. So, if video recording X.avi would be stored in directory *Video/A/B/C/D/X.avi*, then the associated audio file X.wav should be stored in *Audio/A/B/C/D/X.wav*, the EEG in *EEG/A/B/C/D/X.bdf*, the transliteration in *Translit/A/B/C/D/X.txt*, and the annotations in *Annotations/A/B/C/D/X.TextGrid*. Note that each directory contains files of uniform type. This greatly simplifies analysis, versioning, and backup. This scheme works best when files referring to the same recording file share the same name, e.g., like the *X* in the example.

2.3 File names

File names are tricky. It is best to ensure that every filename in the corpus is unique. If not, data will get lost if a file accidentally ends up in the wrong directory. When working with existing non-unique filenames in a more complicated corpus, there is an easy way to make them unique: Prepend the directory path in front of the filename. For example, if there are many files with duplicate names in different subdirectories, e.g., *Audio/A/B/C/D1/X.wav* and *Audio/A/B/C/D2/X.wav*, converting the filenames form *X.wav* to *A_B_C_D1_X.wav* and *A_B_C_D2_X.wav* would suffice to make every filename unique. Such a change can easily be scripted and automated in, e.g., *Praat* [18]. Note that using spaces in file names can seriously complicate automatic processing. It is best to use a character like '_' instead of a space.

It is also very convenient when the filename transparently indicates where it belongs and what it contains. For the example of a two camera recording in a dialogue video corpus, file names could start with *DVA[num][FM][age][A-Z]* for recordings of camera A and *DVB[num][FM][age][A-Z]* for the recordings of camera B. The *number* would be the number of the dialogue, *[FM]* would indicate Female or Male speaker, the age would give the age in years of the speaker, and the letter *[A-Z]* would indicate the speaker in view. So, any file whose name starts with *DVA14M62W* would be from camera *A*, dialogue *14*, and male, 62 year old speaker *W*. The corresponding partner recording in this would be *DVB14M72X*. The file type tells us what is stored in the file. *DVB14M72X*.dv for the uncompressed video, *DVB14M72X*.avi for the compressed video, *DVB14M72X*.wav for the audio. *DVB14M72X*.TextGrid for the annotations, etc. [13, 14]. Note that the above practice of prepending directory names in front of the filename to make the filename unique also makes it transparent where the file belongs in the corpus.

It is common to annotate or extract parts of the items in a language corpus, e.g., sentences and words from texts, or utterances and words from spoken language recordings. When such parts have to be named, it is very useful to prepend the item name with the (unique) name of the originating file. For example, finding an utterance recording with the name *DVA14M62W1J* (e.g., turn *1*, sentence *J* of recording *DVA14M62W*), it is easy to find the relevant (meta-) information in the corpus. This transparency is also useful for checking whether selections and (meta-) data are correct. Some quite elaborate item naming schemes can be found in [12, 13, 14].

3 Content and documentation

The contents of a corpus can be divided into three categories:

- *Primary data*
  Original recordings and observations
- *Meta-data and documentation*
  Information on Primary data
- *Derived content*
  Everything that can be reproduced from the other two

The first two, *primary* and *meta-data* are the real content of the corpus. The third category, *derived content*, is only stored because of convenience. The fact that *derived content* can be reproduced does not mean that it should be reproduced every time. Only that when *primary* and *meta-data* are changed or adapted, the dependent *derived content* can be regenerated. In storage and back-up procedures, *primary* and *meta-data* should be handled with extra care. *Derived content* often does not have to be backed up at all.

It must be emphasized that derived content should preferably be generated by automated methods, i.e., scripts. Manual labor is very difficult to reproduce. If an original recording is split up by hand into smaller parts, e.g., utterances, it will be very time consuming to generate a new set after even a minor change. However, if the original segmentation had been stored in a TextGrid file, then it would take only a few changes in the TextGrid annotations and running a script to regenerate the changed set. Moreover, anyone can check whether the original segmentation was indeed correct.

This can be generalized to other aspects of the corpus. Whenever possible, construction and maintenance of a corpus should be automated. Preferably with scripts or other documented means. These automated procedures should be organized in such a way that users of the corpus can maintain or reconstruct the corpus using as little external (insider) knowledge as possible.

3.1 Participants and procedures

In a language corpus, it is important to store all relevant data of the speakers, authors, and other subjects that participated in constructing the corpus. Up front, it is often difficult to decide what is relevant and what is not. So there might be an incentive to include as much information as possible. However, many pieces of information about subjects are privacy sensitive and should not be distributed or even included in the corpus. Such privacy sensitive material should be handled with extra care. Some personal data is considered so sensitive in some jurisdictions that it is illegal to collect and store them. Here we can give only some rules of thumb. Please, inform yourself about the laws in your jurisdiction.

Sensitive information should stay “in house” and not be distributed without a signed informed consent from the subjects. *Highly sensitive* data should only be stored in a secure environment. An easy way to prevent mishaps is to encode all subject names at the earliest possible moment and store (highly) sensitive data, like contact information, off-line or printed on paper, in a locked closet. At a later stage, relevant information can be compiled from the offline storage and anonymized for the corpus. There are special guidelines for working with data from children and medical records. The short version is that you should not share such data and keep everything behind locks or in a secure environment. The long version is that if you work with such data, you might want to re-read the relevant guidelines.

Every study will have its own requirements for data about the subjects and language. Often some aspects of language use and the subjects are not relevant for the original research. However, other users of the data might need such personal data. So it often pays off to record them anyway. Some types of data are almost always relevant. Data of subjects and circumstances that are generally relevant to construction and use of spoken language corpora can be listed as:

- Contact information of the subjects
  This is highly sensitive and should not be shared
- Age in years
  Date of birth is sensitive
- Sex/gender
- Language variant used, native language, other languages
  Place of origin, e.g., postal code (this can be sensitive if too specific)
- Hearing and speaking problems (mostly, the absence thereof)
  When relevant, the nature of any pathologies (highly sensitive)
- Recording: Date and Location of the recording and the Name of the person doing the recording
- Equipment and recorder settings

Some data, like the language or technical details of recordings and procedures, tend to be fixed in which case they can be stored in the general documentation.

It is very important to have all the subjects that participate in the creation of the corpus sign the relevant documents (see section 4). Most importantly are copyright forms that transfers all copyrights to the corpus maintainers and informed consents for speakers and experimental subjects. Make sure that these forms and declarations explicitly include a reference to the distribution of the materials [12, 13, 14].

3.2 File formats

In general, you should do as little conversions between file formats as possible. But if the aim of the corpus is to make the data available to outsiders, then it makes sense to chose file formats that are in common use. In general, plain text and all formats read by *Praat* [18] should be fine. Some applications, like office or specialist video software, produce files that are difficult to use without the exact same software or even computer platform. In such cases, it is best to add copies in generally readable format. Note that, at the time of writing, video codecs are a complete mess [19].

If possible, for non-generally readable file formats, ensure that copies using the the following formats are also available in the corpus:

- Word processor, e.g., *MS Word*: Plain text and PDF
- Spreadsheet and database, e.g., *MS Excel*, *MS Access*: Export as Tab Separated Values (.tsv) or Comma Separated Values (.csv)
- Audio: Uncompressed WAV files, else anything written by *Praat* [18] is good
- Pictures: PNG or JPEG files
- Video: Compress as little as possible, original DV or MJPEG would be nice, but at least use something that can easily be played with *VLC* [20] or *ELAN* [21] (note, *ELAN* video support is platform dependent)
- Annotations: *Praat* [18] TextGrid files or *ELAN* [21] EAF annotations

3.3 Annotations

Annotations are interpreted here as texts that are (time) aligned to the recordings [22]. These notes will discuss annotation files like those used in *Praat* [18] TextGrids and *ELAN* [21] EAF. An annotation generally has a start and end time, a *Tier*, and a text. Other set ups are possible, c.f., [22, 9, 10]. In corpus construction and maintenance, annotations have two independent functions. The first is to segment the recordings so relevant fragments, e.g., utterances, turns, or words, can be identified and retrieved. The second is to add additional information to the recordings. In general, annotations include a transliteration of the speech as text and a segmentation in utterances (or Inter-Pausal-Units), turns, and sometimes phrases or words. Other annotations used are (in random order): Gaze direction, word stress, backchannel utterances, disfluencies, gestures, or emotions. For adding any information related to fragments of the speech or language, standard annotation files should be used wherever possible. These should be stored as text files (*short text file* in *Praat* [18]).

It is most efficient to use a segmentation stored in annotation files to split up recordings. It is fairly straightforward to code a *Praat* script that takes a TextGrid and copies out selected intervals for, e.g., listening experiments. This annotation file is then instant documentation of the segmentation and selection. The segmentation can also be easily adapted and recreated if needed.

3.4 Metadata and CMDI

Corpus data have only little value without information about the speakers, authors, task, and recording circumstances. This type of data is called *metadata*. In general, each item in a corpus has its own metadata. If metadata is available, it should be stored in its own, parallel directory tree, just like the annotations. The top level can be called *Info* or *Metadata*. Please consult the relevant standards documents when using browsable metadata hierarchies, [23, 24].

There is extensive literature about the use of metadata in corpus construction, e.g., [25, 26, 27, 28, 29]. There are several international metadata standards for language corpora. Early ones are *TEI* [30] for text and *IMDI* [23] for multi media data. The *IMDI* standard has been adapted inside the Clarin project [31] to the *CMDI* standard [24, 32]. The advise is to use *CMDI* to code browsable metadata.

How to compile metadata is outside the scope of these notes. Readers can consult the Clarin user guide by Wittenburg and van Uytvanck [25] for more information as well as the relevant corpus handbooks [1, 2, 3]. It is important to remember the recommendations from [25]:

- Always start as early as possible to collect and create metadata. Otherwise chances are high that information is lost or that fixing the incomplete metadata records afterwards will be very costly.
- Try to achieve a high but reasonable level of granularity.
- Try to reuse as much as possible. It should save you work and will enhance the interoperability.
- Be aware that there are conversion methods in place for the most widely used formats - there is no need to reinvent the wheel.

3.5 Scripts, programs, and applications

As much as possible the construction of the corpus and the analysis of the data should be automatized. This even holds for the statistical analysis and the production of figures for publication. Automatization is mostly done through the use of *scripts* or other programs. Experience shows that quite often, such scripts have to be consulted later to (re-)verify the validity of data and procedures, adapt the corpus or analysis to new requirements, and to reuse them for new tasks. Such scripts should be an integral part of the corpus. They will be necessary, not only to manage or rebuild the corpus, but also to understand the corpus and the analysis of the data. If special purpose programs have been used, it should be considered whether the version used in building or analyzing the corpus can be included with the corpus.

By adding the scripts in fixed directories, either in a separate *Scripts* tree of sub-directories, or as part of the *Documentation* sub-directory, it will be possible to use relative file paths. Relative file paths are necessary to allow the corpus to be moved from one computer to another. If at all possible, relative file paths should be used in scripts. See the *Praat* manual for examples of the use of relative file paths in scripts [33].

3.6 Documentation

Any information about the corpus that is not stored with the annotations or metadata is documentation. The aim of the documentation is to help users to understand the contents of the corpus, maintainers to extend and correct the corpus, and others to recreate a similar corpus. The documentation is also an archive for what has been done. This will be needed when data from the corpus is used for a publication. So the documentation should at least contain all information that is required for publishing a scientific paper about the contents of the corpus. The documentation section is also a good place to store information about the corpus itself, e.g., contact and license information, scripts and other special purpose software, errata, literature references, manuals. It is often convenient to store compilations of global metadata in the Documentation section, like speaker data and recording lists.

In a corpus, just create a sub-directory called *Documentation* and store any useful files in there. This directory can contain sub-directories as is convenient. Obvious content of a *Documentation* sub-directory are:

- Contact information, copyright and license documents of the corpus
- A changes and errata list
- A diagram of the corpus structure
- Anonymized lists of subjects: Speakers, listeners, and other experimental subjects
- Relevant parts of anonymized information about the subjects
- A list of recordings, e.g., with dates and the names/codes of those involved
- All technical details of the recordings or experiments, preferably with photographs
- Copies of any documents and forms given to the subjects, e.g., copyright forms and informed consent declarations (but *not* copies of the signed documents)
- Copies of any documents and manuals used during the construction, recording, and annotation of the corpus
- All scripts and special purpose programs used in the construction and analysis of the corpus (when not stored in a separate subdirectory)
- All publications based on the corpus
- A bibliography of the relevant literature, e.g., as a BibTeX reference file

3.7 Backup

The advised mind-set is to imagine, every day, that the building where your data reside burns down to the ground. Then think of how you would want to continue with your work. On regular moments, you should try out whether you really *can* continue your work after the building has burned down to the ground.

3.8 Version control and repositories

A spoken language corpus generally consists of static language and speech data and a monotonically increasing amount of annotations, scripts, and other textual materials. The static language data can be backed up once and then the backup has to be updated only infrequently. However, it is almost always good to keep track of changes in the textual materials in a fine grained time scale. Errors happen and often changes will have to be undone. For this, versioning systems should be used [34]. In short, a version control system will store "snapshots" of your texts. It allows you to roll-back your system to any point in its history and even to pick and "undo" individual changes made to the system. The most important systems allow to merge changes made by different maintainers at different times. Note that version control systems are mainly useful for textual data, e.g., annotations and documentation. Most systems cannot well handle binary data, like pictures and word processor files.

A version control system will store the history of a project in a *repository*. Such a repository can easily be made available from a website or project server. Such repositories are excellent means to distribute and update corpora. A popular choice of version control system is *Git* [34]. See the slideshow at [35] for a short introduction. Explaining the use of version control systems is outside the scope of these notes. But there are excellent tutorials and manuals for most popular system, see the links and references in [34].

4 Distribution

The aim of a language corpus is giving other people access to language data. While designing and constructing a corpus, the ways the corpus will be accessed have to be taken into account. The corpora discussed here are small enough to distribute in full. Therefore, there is no pressing need to consider elaborate online search and selection solutions. For practical reasons, subsets of the corpus can be made "pre-canned" for download. Otherwise, it is simplest to allow wholesale copying of the corpus to prospective users. The simplest technical solution is to install a web server, e.g., *Apache*, and point it to an index file in the top directory of the corpus. The technological details of this are beyond the scope of these notes, but you can look at the *Apache* documentation for more information [36].

Beyond the file format compatibility issues and useful documentation, there are a few legal matters that have to be dealt with before a corpus can be distributed.

4.1 Copyrights

In general, everything spoken or written will fall under copyright protection. In a language corpus, the language that are the primary data in the corpus are almost always protected by copyright law. When spoken language is involved, the speakers have a comparable right as "performers". So have the editors of the material. All the written materials and documentation will be protected under copyright law. This list could be extended ad infinitum. Under copyright law, anyone who wants to copy or distribute protected works. a corpus, needs written permission of the "owners" of the copyright. That would mean all those involved in constructing the corpus, which is impractical to say the least.

The solution is to ask everyone who participates to sign a copyright transfer form. In this form, the participant transfers all copyrights to the "owner" of the corpus, who will then be the sole owner of copyright. This new owner is some legal entity, that will manage and distribute the corpus. In many cases this entity will be some part of the university or research institute. However, it can be the creator of the corpus, or some non-profit organization, e.g., Nederlandse Taalunie. Just remember that this new owner will be the legal owner of the corpus. This entity will decide what will happen with the corpus. So some care might be taken to chose a suitable entity to transfer the copyrights to and to make good, binding agreements about the future of the corpus.

The question on who should all sign a copyright transfer form is not easy to answer. It is best to be "inclusive" and just ask every person who touches the data to sign a copyright form. If in doubt, ask for a signature. This policy works best when people give their signature *before* they start with their contribution. This includes all subjects who speak. However, there could be problems when spontaneous, or unscripted, speech is recorded. Then the speaker would have signed a copyright transfer before she or he knew what was said. In such cases it is best to confirm the signature again after the recording. That is, when recording unscripted language use, ask the speakers to sign the forms a second time after the recording. Sometimes it is prudent to give the speakers a copy of the recording and offer them the option to retract their consent. For an extended discussion of this topic, see [13, 14].

Drawing up a copyright transfer form should be done by a specialist in copyright law. In most cases, this is too much work (and too expensive). It is then best to take a boilerplate form and adapt it to the needs of the corpus. It must be stressed that the transfer forms should make clear, upfront, to subjects how the recordings and the personal data might be used. In practice, this means that the different options, e.g., publishing recordings and meta data on the internet, have to be written explicitly into the copyright transfer forms. A good guide seems to be that corpus creators are specific about the intended uses whenever possible. At the same time, an effort should be made to be inclusive and prepare for potential, future, uses by yourself and others. All the "legal" information has to be made available also in layman's terms in an informed consent declaration (see below). Obviously, subjects should have ample opportunity to ask questions about the procedures and use of the recordings. An example copyright transfer form can be found at the IFA Dialog Video corpus [15].

4.2 Informed consent

Having experimental subjects, or speakers, sign legal documents does not imply that they fully understand the effects their participation can have on their lives. However, it is paramount that all subjects fully understand the potential consequences of participating. To ensure that every subject has understood and accepted the (potential) consequences of their participation, they should sign an *informed consent* document, e.g., [37].

Informed consent is a *process* to enable a subject to make an *enlightened decision* whether to participate in a study or not (*Nuremberg Code, 1947*). The informed consent document should explain in plain and easy to understand language what will be expected from the participants, and what the consequences can be of their participation. This informed consent document should also state what will happen with their contributions. For example, if spontaneous dialogue is recorded and will be published on-line, it should be ensured that the speakers know about this. If such a publication on-line might possibly affect their future life or career prospects, this should also be made clear to the subjects (and ways should be found to prevent or remediate such outcomes). There are extensive regulations about how to execute the informed consent procedure and how to draw up the documentation and forms. The exact rules differ per country and might even differ per institute ethical committee. The rules as they are described in the Good Clinical Practice (GCP) guidelines are a good starting point as they cover the universal base [38].

Journals and conference publishers generally require informed consents from all subjects whose data are used in a publication. Special consent is required when pictures or video clips of subjects are used in a publication or presentation. In practice, *both* signed copyright transfer forms *and* informed consent documents are required before any data in a corpus can be used. And it cannot be stressed enough that both documents should contain clauses about *all* intended and possible future uses of the data.

4.3 Privacy

It is generally accepted that researchers have a *duty of confidentiality* with respect to informants and experimental subjects and should respect their privacy [1], e.g., [37]. Only in exceptional cases will the names or other identifying information of experimental subjects be revealed. This rule is embedded in the law and there are almost no exceptions when, e.g., minors or patients are involved. When co-authors or colleagues participate in experiments, it is common to use their initials in publications. However, names of external subjects should always be securely coded and initials are not considered secure. For an individual paper, enumerating subjects, e.g., S_1-S_i, often suffices. In a corpus, this can become unwieldy, especially in more complicated corpora containing multiple contribution from individual subjects, sometimes in different roles. Unless ethical rules require total anonymization, each subject should get an individual, fixed code, a token, that is valid for the complete corpus. It is obvious that the real names and contact information should never be stored with the corpus.

If at all possible, subjects should get a (random) code at enrollment. Trying to fix the internal codes after recordings of experiments are completed is fraught with problems and frequently leads to persistent errors. The easiest system for subject coding is simply giving numbers or letters in order of enrollment. This can be $[S1\ldots]$, or $[A, \cdots, ZZ]$. It does not matter for such a procedure that some subjects will drop out. In special cases, it can be necessary to adopt more involved protocols. Needless to say that the decoding lists that link subject codes to identities and contact information should be stored and backed up securely.

An example of a special case was the use of patients enrolled in long term follow up research. Many patients participated in several studies while it was not always clear which patients had already participated before. Unique subject codes were constructed by encrypting the unique hospital patient ID (using a password). This ensured that every patient's contribution to every study would be labelled with the same identifier token even when the previous contributions were not known. The encryption procedure was designed to make it impossible to extract personal information from the codes. Cryptographic protocols are very fragile and can easily fail. Caution must be exercised when using them.

A few suggestions for handling subject privacy:

- Keep all personal and contact information securely off-site, make sure there is a printed paper back up at a secure site
- Remember that date-of-birth, zip codes, and IP addresses are sensitive information
- Assign each subject a unique anonymous code at enrollment and *never* use a name or initials
- Subject codes should be sequential, random, or if that is not possible, cryptographically secure
- Use subject codes for *all* references to subjects, internal and external
- Do not publish recognizable audio, pictures, or video without the explicit written consent of the subjects

[1] The legal aspects of this subject are discussed below

- If sensitive information *has* to be shared with outsiders, require a signed, legally binding, *promise of confidentiality* [37] or Non Disclosure Agreement (NDA).

4.4 Licenses and moratoria

As discussed in section 4.1, the owner of the copyrights to the corpus must give written permission for use of any part of the corpus. This can be on a case-by-case basis, which is impractical, or more efficiently by way of a copyright license. A copyright license is a written permission to copy and distribute work under copyright, i.c., a corpus. A copyright license determines how a corpus can be used and by whom, and what can be done with the results. These notes will only discuss the *Open Data* case [39], where a corpus is shared on liberal terms.

There are two families of copyright licenses relevant for *Open Data* compatible corpora, *Free and Open Source* licenses for code and software [40, 41], and *Creative Commons* licenses [42] for all other materials. When choosing a license, it is strongly advice to adopt an existing license. In practice, adopting a newly written license has only downsides. Lists of popular licenses can be found at:

- *Creative Commons* [42]: `http://creativecommons.org/licenses/`
- *Free and Open Source* [41]: `http://opensource.org/licenses`

It is often impractical for a researcher to wait until her project is completed before adding her data to an *Open Data* corpus. For logistic reasons, it might be even preferable to store all primary data directly in an existing corpus. In such cases, the researchers would not allow distribution of their data before they have finished their primary analysis and publication. This is handled by a *moratorium* on the data. When the data are added to the corpus, the "license" specifies an agreed date, after which the data will be available to the "public". Alternatively, the data will be made available after the official publication of a certain paper. When constructing a corpus, the possible inclusion of such moratoria should be considered.

5 The GDPR

Disclaimer: I am not a lawyer and this is not legal advice. Please consult your lawyer for legal advice about your project.

The European General Data Protection Regulation (GDPR) rewards its own consideration. The GDPR is a comprehensive legal framework that governs the collection and processing of information about natural persons in the EU [43]. The GDPR covers all collection and processing of data from EU citizens by anyone who offers goods or services, or who monitors EU data subjects, irrespective of where the data are stored or processed [44]. It comes into force 25 May 2018. Some countries have already implemented parts of the GDPR well before that date, e.g., the Netherlands. Its most reported aspect is a testimony to the determination of its authors to make even the largest of companies comply. The maximum fines are set to 20 million euros or 4% of global turnover, whichever is larger. However, the goal of the GDPR is mostly to harmonize the EU privacy rules to foster market innovation while at the same time protecting the citizens' privacy from corporate abuse. To that latter end, anyone who collects or processes personal data should put the privacy interests of the subjects covered by these data first.

The main concerns of the GDPR are to create a EU wide, uniform market space for information and changing the ways privacy issues are handled in non-research settings, e.g., commercial companies and public bodies. The inconveniences for science and research can mostly be seen as collateral damage. Several exceptions are made in the GDPR for research. However, many of these exceptions do *not* cover health related research. And health related research is defined very broadly as anything that can conceivably be used to assess or influence the health status of natural persons. One relevant research exception of the GDPR covers secondary use, the use of existing data for different purposes, without having to obtain consent nor having to limit the duration of the use. With respect to shared speech and language corpora, it is not advisable to rely on such a secondary use exception. It is best to obtain informed consent from each subject for inclusion in a database, unless there are exceptionally pressing reasons and there is good legal advice.

An important new focus of the GDPR is *accountability*. It is not enough to *be* in compliance, anyone who handles, collects, stores, or processes personal identifiable information (data controllers) must *demonstrate* how they are in compliance with the GDPR [44, 45, 46]. This means that documentation must be maintained about all measures taken and procedures installed to ensure compliance by the owner of the data, i.e., the controller, and any other party processing the data.

Before data can be collected or disseminated, the ethical and legal aspects must be evaluated. It must be clear what are the risks (to the subject) and benefits (to society) of the research and corpus.
In general, before collecting or disseminating corpus data the following points have to be considered:

- Privacy Impact Assessment (PIA)
- Privacy by Design technology
- Approval from the Medical/Research Ethical Committee (M/REC)
- Consultation and/or approval from the relevant Data Protection Officer (DPO) or Privacy Officer (PO)
- Explicit, written informed consent by the data subjects, and copyright transfer
- Data Transfer Agreements (DTA), Non Disclosure Agreements (NDA), Promise of Confidentiality (PoC)

5.1 Privacy Impact Assessment (PIA)

To be able to justify collecting and processing personal identifiable information (PII), it is important that the likely risks and benefits to the subject are known and that efforts are made to minimize these risks. A *Privacy Impact Assessment*, or *Data Protection Impact Assessment* (DPIA) in the GDPR, is defined as ”a process which assists organizations in identifying and minimizing the privacy risks of new projects or policies” [47]. This means that all the possible consequences for the privacy of the subjects that might result from collecting, processing, and disseminating the intended data should be discovered as well as all the ways any negative consequences can be eliminated or minimized. An official PIA is mandatory, when ”the processing is likely to result in a high risk to the rights and freedoms of natural persons” [45]. However, running a critical assessment of the expected risks and benefits of the proposed project would be a good idea irrespective of the formal requirements of the law. It will deliver information that will help with the design of the informed consent forms and the corpus itself and will also help the ethical committees and the data protection officer to evaluate the proposal.

How best to do a PIA is still an area of ongoing research [48]. Not surprisingly, there are yet no generally applicable frameworks to do a PIA in speech and language research. However, when creating a speech or language corpus, it will often suffice to think hard about what data to include and how to minimize the probability of unintended consequences. With this information, the ethical committee and data or privacy officers can better evaluate your plans and possibly point out areas needing improvement. It is very likely that these institutions will even demand such an assessment from you before they will even agree to evaluate your proposal. A well documented PIA will also play an important part in demonstrating compliance with the GDPR [45]. Being able to demonstrate compliance is a prerequisite for being allowed to handle PII.

Note that the PIA is a ”living document” and should be updated whenever new technologies or information become available [48, 49]. The PIA must certainly be updated when there is a change in the way data are handled that can affect the privacy risks. If the risks change, policies might have to be changed and possibly. In extreme cases, it might be necessary to contact the subjects about changed circumstances or try to obtain new informed consent.

5.2 Privacy by Design technology

Data security is hard, very hard. Therefore, the GDPR has build-in incentives to use technology that has security build-in from the start, what is called *Privacy by Design* [46, 50, 44, 51]. Privacy by Design starts with keeping information out of the corpus. The GDPR is only concerned with PII, i.e., information that can be linked to a natural person. Privacy by Design aims to prevent that stored information can be linked back to a natural person by unauthorized parties. That starts by preventing data breeches and data loss. What is not stored, cannot be lost or leaked, so the first line of defense is to keep the most obvious PII out of the corpus. There are direct identifiers, e.g., all contact information, and indirect identifiers, information that allows to differentiate a person from others, e.g., IP addresses. But it is well known that almost any information can be used, in combination with other facts, to identify a person, e.g., birthday, zip code, and birthplace [52]. For instance, some innocuous time and location information is generally enough to identify almost anyone [53].

By keeping information that is not needed out of the corpus, the possibilities for and severity of data loss and breeches are reduced.
Four methods are mentioned by name in the GDPR: Data minimization, Anonymization, pseudonymization, and encryption.

Data minimization Obviously, what is not there cannot be lost or exposed. One of the prime objectives of privacy by design is to reduce the amount of data stored at all. The first approach is simply removing all data that is not necessary for the task at hand. All data that is left should be examined for precision. Whenever precision is not needed, remove it. Date of birth is a known privacy risk and almost never needed. Replace it with *age*. For young children, the age in months might be appropriate. For older subjects, decadal age brackets might be enough. For ages above 80, subjects might be pooled. Full zip codes are a privacy risk. Truncate them, or use broad regional areas.

A special case are images, e.g., pictures, movies, MRI. These tend to contain metadata that should be stripped. If possible, add censor bars to cover the eyes in pictures and movies and use equivalent practices for MRI. Note that "ear prints" are like finger prints and the outer ear should be masked in MRI's too.

Anonymization After anonymization, it is impossible to link a data item to a person. Anonymous information falls outside the scope of the GDPR and can be used freely. This would make anonymization an attractive option for using information. There is a large body of research on the limits of anonymization [54, 55]. Such research has shown that guidelines on de-identification published in the USA (HIPAA [56]) and practices proposed in the UK (the Care.data initiative [57]) are inadequate to protect the identity of the subjects to the level of security that was intended. Under the GDPR, these "de-identification" practices are not considered anonymization or possibly not even good pseudonymization. All in all, the use of anonymized data in research suffers from security problems and is fraught with uncertainty and ethical questions [58].

In the current context, the results of the research on anonymization can be summarized as: If data is useful, it is not anonymous and if data is anonymous, it is not useful.

Pseudonymization With anonymization an elusive goal, the next best option is pseudonymization. Pseudonymization, or key-coding, is already the accepted norm in all research involving humans. Therefore, there is little that the GDPR will change in this respect. But pseudonymization is important as "good" pseudonymization is necessary for any research exemptions under the GDPR, derogations in GDPR parlance. Good pseudonymization should have no link between the personal identifiers it protects and the code used to identify the records. This means that, e.g., initials are *not* good pseudonymization codes. Unencrypted hashes or message digests of personal information, e.g., names, email addresses, or patient id's, are not adequate as they can easily be broken. Simply run a list of all known names, email addresses, or patient id's through the hashing algorithm and pick the matching hash. However, using random codes or encrypted hashes are considered good pseudonymization. Do keep in mind that using encryption is tricky.

Encryption The use of encryption is mentioned in the GDPR as one of the components of *Privacy by Design*. Encryption can be seen as the digital equivalent of "under lock and key". In practice, it means that all privacy sensitive data should be stored and transmitted in encrypted form. And as with real locks and keys, sensible key management procedures should be in place, and enforced. As encryption is explicitly mentioned in the GDPR, there is no excuse for *not* encrypting all data when not in direct use. With respect to the use of encryption, the only sensible advice is to consult professionals.

As a minimum, projects should use full disk encryption on *all* computers, laptops, and computer storage used. This means, *Bitlocker* on Windows, *Filefault* on Mac OSX, and *LUKS* on Linux. A problem of these solutions is that they are not practical for data exchange between computer platforms. For all practical purposes, *Bitlocker* and *Filefault* volumes are only readable on, respectively, Windows and Apple computers. *LUKS* is rather impractical to use outside of Linux computers. There are multi-platform alternatives that do run on all three computer operating systems, like the TrueCrypt successor *VeraCrypt* (`https://www.veracrypt.fr`). Most likely, your institution will have a institution wide policy with regard to using encryption in computers (if not, they should have).

It is important to realize that it is very likely a breach of the GDPR when privacy sensitive data are stored in any kind of public cloud storage. Projects working with privacy sensitive data should refrain from using *any* kind of cloud storage for *any* project data unless they get clearance by the responsible Data Protection

Officer. That is, no Google Drive, Dropbox, OneDrive, or iCloud, for project data. This also holds for email communication that might (accidentally) contain sensitive data. Data should be securely encrypted whenever it is communicated or exchanged. Projects should also consider the use of approved secure email or message services if at all possible. In this matter, always consult with the Data Protection or Privacy Officer responsible.

Repository direct access-analysis A different option is to not distribute the data to be analyzed at all, but to bring the analysis to the data [59]. Third parties that would like to analyze the data could specify the desired analysis in a known format. The owner of the data would then perform the analysis and return the results. If the results cannot be re-identified, privacy is preserved. This way, no one outside of the owner has access to PII. Examples of speech research would be to perform an extensive voice or spectral analysis followed by statistical modelling, or to train a speech recognizer from the data. In most cases, the results would not allow re-identification of the data subjects. The DataSHIELD Open Source project implements such a system in the field of epidemiology and genomics research [60, 61, 62, 59], the COINSTAC project does this for processing brain image data [63]. ViPAR [64] is developed as a more general platform for repository direct access analysis and supports a broader array of functions.

A *repository direct access analysis* can be envisaged as a kind of web-service where clients give the parameters of an analysis and the owner of the data runs the analysis. The client only receives the outcomes of the analysis. Precautions must be taken to prevent disclosures that might identify individuals, e.g., outliers, extreme values, and scatterplots are generally not shared. This service could be extended by running certified client software on the data inside "sandboxes" to, e.g., for deep learning and training classifiers and recognizers. After vetting the results for privacy risks, the results can then be disclosed and the sandboxed application can be destroyed. Properly implemented, such a system would obviate most privacy concerns.

5.3 Approval from the Medical/Research Ethical Committee (MEC/REC) and the relevant Data Protection Officer (DPO) or Privacy Officer (PO)

The collection and use of data for academic research requires approval of a Research Ethical Committee. When patients are involved, approval of a Medical Ethical Committee is required. These committees evaluate the risks and benefits of the proposed project and decide whether the benefits to society outweigh the risks to the subjects. Every institution will have their own procedures and requirements for this process. These committees will most certainly need to see all the legal paperwork, e.g., the Informed Consent and copyright transfer forms as well as all the data transfer agreements, non-disclosure agreements, and promises of confidentiality that will be used.

To judge whether a proposed project strikes the right balance between risks and benefits, data about the expected risks and benefits are needed. The PIA and technical measures discussed above will help to evaluate the likely risks and benefits.

When something goes wrong and the corpus is somehow in breach of the GDPR, the owner of the corpus or the institution that created it, or both, will be held responsible. Naturally, those who bear the consequences of anything going wrong will have a decisive voice in determining whether and how the corpus will be constructed and distributed. Most big institutions will have a data protection officer or privacy officer, or a another person who is responsible for cleaning up after a data breach. These officers should be consulted before a corpus is constructed. Better, they should be consulted even before a corpus is designed. Most likely, their approval is needed before the start of the work.

What is also needed in both evaluations is a data management plan [65]. This plan gives detailed information on what data will be stored, for how long, and for what use, how it will be stored and disseminated, how it will be secured, and who are responsible. It will also describe the access and use policies and the procedures for consent retraction, if relevant, and breach notification. In short, everything necessary to show that the corpus is constructed and used in compliance with the GDPR.

5.4 Explicit, written informed consent by the data subjects, and copyright transfer

Central in the whole privacy discussions and regulation is the informed consent by the data subject. Data can be collected and processed if, and only if, the subject has given an explicit consent to the collection an processing. In the research context, this constitutes no change from the accepted standards. All research involving humans already has to be based on explicit informed consent well before the GDPR came into view.

In clinical research this has been codified, in, e.g., the Good Clinical Practise guidelines [38]. Although all attention is focused on the GDPR, informed consent in a health care context is governed by the Clinical Trial Directive (CTD) and the upcoming Clinical Trials Regulation (CTR) [66, 67]. As health care is a concern of the member states of the EU, the consent rules are set by the member states too. This means that the rules vary across borders, complicating international collaborations and data sharing.

Informed consent is not the signature on a form, but it is a process to enable a subject to make an *enlightened decision* to participate or not (*Nuremberg Code, 1947*). The GDPR states that consent can always be withdrawn, even retroactively by giving a limited *Right to be Forgotten*. A data subject can withdraw their consent at any moment for any reason. Data subjects can also ask for their data to be removed. Here there is a difference with the CTR, which too allows subjects to withdraw their consent at any moment, but does not require that data collected before withdrawal of the consent has to be removed. It is not clear whether these consent practices will converge.

Under the GDPR, consent must be specific. The use of the data and the time the data will be stored must be indicated in the informed consent. Any new use requires a new informed consent from the data subject. This too clashes with the CTR which allows more open ended nature of a one-time consent in medical research [66]. It is obvious that requesting new consent for every new research question would be a burden on the data subjects and make reuse of data impractical and uneconomical. The differences between the approach of the GDPR and the CTR leads to ambiguity and uncertainty in the collection and use of data [68]. How specific must an informed consent be? When have subjects the right to have their data removed from a project? What information must data subjects be given about the uses and users of their data?

The subject of copyright transfers has been discussed in section 4.1. Note that the copyright transfer and its consequences also have to be part of the informed consent.

5.5 Data Transfer Agreements (DTA), Non Disclosure Agreements (NDA), Promise of Confidentiality (PoC)

In the ideal case, the corpus contains only data that can be shared without restrictions because the subjects consented to full publication and both the ethical committees and data protection officers approved the sharing of the data. However, this is often not the case. Especially if the corpus contains health related information, e.g., pathological speech, the odds are that some or all of the information will have to carry restrictions on use. In practice, this means that anyone who wants to obtain and use the data will have to sign legally binding documents in which they agree to not disclose the data to others, to abstain from re-identifying the subjects, and to not divulge any information that might help others to re-identify the subjects. Users must also agree to destroy all copies of the data after a certain time limit or if they have been found in breech of the agreement. Those who obtain copies of the data must also agree to apply the required level of data security and notify the publisher (controller) of the data and the authorities of any security breeches in relation to the data.

It might be prudent to investigate whether it is possible to include provisions in the agreements that require that the data must be destroyed in case of an imminent risk of re-identification of the subjects and that data from individual subjects must be deleted upon request (Right to be forgotten).

Depending on the nature of the sensitive data that have to be protected and the relation with the user of the data, different legal instruments can be used: Data Transfer Agreements (DTA), Non Disclosure Agreements (NDA), Promise of Confidentiality (PoC). This choice can best be made in consultation with a legal department with experience in this matter. Writing such papers is work for legal specialists. The legal department that chooses the instrument will also know who should draw up these legal papers.

6 Concluding remarks

In theory, anything should be possible with the right *Informed Consent*. However, in practice the ethical committees will limit what can be asked from subjects. The rules say that a valid Informed consent must be specific and cannot be open ended. The wording of this rule will vary between EU member states. On the other hand, the principle of self-determination and the autonomy of the patient allow patients to make any data about their health status public, as TV-reality shows illustrate almost daily. It is currently unknown how this discrepancy between self-determination and autonomy of the patient and legal protection under the GDPR and CTR will work out.

Irrespective of the way the questions about the limits of patient informed consent will be resolved, it is clear that those who want to publish or use corpora holding privacy sensitive data must become “competent” in the securing and protection of such data. Data should only be shared with partners and users that themselves have shown to be competent in handling sensitive data. Professional organizations and associations should strive to compile guidelines for handling research data in their field of expertise, e.g., like the *GCP* does for clinical tests. Such guidelines can become the focal points of binding Codes of Conduct (CoC) and Certification. Such binding CoC and certifications are the preferred way for streamline processing of sensitive data under the GDPR.

In many cases, it will not legally be possible to share data liberally. A promising solution for these cases is to bring the analysis to the data, instead of the data to the analysis. That is, the owners of the data set up a service where the analysis is performed on the data, but the researchers that request the analysis only see the outcomes. Such a service can often be satisfactorily secured against unwanted disclosure while still allowing many researchers "anonymous" access to the data.

7 Acknowledgements

This work has been made possible by an unrestricted research grant from Atos Medical (Horby, Sweden).

References

[1] M. Wynne, *Developing linguistic corpora: a guide to good practice.* Oxbow Books, 2005. `http://www.ahds.ac.uk/guides/linguistic-corpora/index.htm`.

[2] D. Gibbon, R. K. Moore, and R. Winski, *Handbook of standards and resources for spoken language systems.* Walter de Gruyter, 1997.

[3] D. Gibbon, I. Mertins, and R. Moore, *Handbook of multimodal and spoken dialogue systems: resources, terminology, and product evaluation.* Springer, 2000.

[4] C. Draxler, "Using a global corpus data model for linguistic and phonetic research," in *Best Practices for Speech Corpora in Linguistic Research Workshop Programme*, p. 51, 2012.

[5] B. MacWhinney, Y. Rose, L. Spektor, and F. Chen, "Best practices in the talkbank framework," in *Best Practices for Speech Corpora in Linguistic Research Workshop Programme*, p. 57, 2012.

[6] C. Cieri and M. Yaeger-Dror, "Toward the harmonization of metadata practice for spoken languages resources," in *Best Practices for Speech Corpora in Linguistic Research Workshop Programme*, p. 61, 2012.

[7] S. Drude, D. Broeder, P. Wittenburg, and H. Sloetjes, "Best practices in the design, creation and dissemination of speech corpora at the language archive," in *Best Practices for Speech Corpora in Linguistic Research Workshop Programme*, 2012.

[8] E. M. Hart, P. Barmby, D. LeBauer, F. Michonneau, S. Mount, P. Mulrooney, T. Poisot, K. H. Woo, N. B. Zimmerman, and J. W. Hollister, "Ten simple rules for digital data storage," *PLoS computational biology*, vol. 12, no. 10, pp. e1005097, doi:10.1371/journal.pcbi.1005097, 2016.

[9] N. Oostdijk, "The spoken dutch corpus. overview and first evaluation.," in *LREC*, 2000.

[10] N. Oostdijk, "The design of the spoken dutch corpus," *Language and Computers*, vol. 36, no. 1, pp. 105–112, 2001.

[11] P. Wittenburg, U. Mosel, and A. Dwyer, "Methods of language documentation in the dobes project.," in *Proceedings of LREC 2002*, 2002.

[12] R. van Son, D. Binnenpoorte, H. van den Heuvel, and L. Pols, "The IFA corpus: a phonemically segmented Dutch Open Source speech database," in *Proceedings of EUROSPEECH 2001 Aalborg*, pp. 2051–2054, 2001.

[13] R. van Son, W. Wesseling, E. Sanders, and H. van den Heuvel, "The IFADV corpus: a free dialog video corpus," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)* (N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, and D. Tapias, eds.), (Marrakech, Morocco), European Language Resources Association (ELRA), may 2008. `http://www.lrec-conf.org/proceedings/lrec2008/`.

[14] R. van Son, W. Wesseling, E. Sanders, and H. van den Heuvel, "Promoting free Dialog Video Corpora: The IFADV Corpus Example," in *Multimodal Corpora* (M. Kipp, J.-C. Martin, P. Paggio, and D. Heylen, eds.), vol. 5509 of *Lecture Notes in Computer Science*, pp. 18–37, doi:10.1007/978–3–642–04793–0_2, Springer Berlin Heidelberg, 2009.

[15] Institute of Phonetic Sciences, Amsterdam, "IFA Spoken Language Corpora." `http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/`, 2013.

[16] Dropbox, "Dropbox." `https://www.dropbox.com/`, 2013.

[17] R. J. J. H. Van Son, "A study of pitch, formant, and spectral estimation errors introduced by three lossy speech compression algorithms," *Acta acustica united with acustica*, vol. 91, no. 4, pp. 771–778, 2005.

[18] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." `http://www.praat.org/`, 1992–2008.

[19] S. Phipps, "Video codecs: The ugly business behind pretty pictures." `http://www.infoworld.com/d/open-source-software/video-codecs-the-ugly-business-behind-pretty-pictures-214525`, 2013.

[20] VideoLAN, "VLC media player." `http://www.videolan.org/`, 2013.

[21] ELAN, "ELAN is a professional tool for the creation of complex annotations on video and audio resources." `http://www.lat-mpi.eu/tools/elan/`, 2002–20013.

[22] S. Bird and M. Liberman, "A formal framework for linguistic annotation," *Speech communication*, vol. 33, no. 1, pp. 23–60, 2001.

[23] IMDI, "ISLE Meta Data Initiative." `http://www.mpi.nl/IMDI/`, 1999–2007.

[24] CLARIN ERIC, "Component Metadata CLARIN ERIC." `http://www.clarin.eu/content/component-metadata`, 2013.

[25] P. Wittenburg and D. van Uytvanck, "Chapter 2. Metadata." `http://media.dwds.de/clarin/userguide/text/metadata.xhtml`, 2013.

[26] L. Burnard, "Metadata for corpus work," in *Developing linguistic corpora: a guide to good practice* (M. Wynne, ed.), Oxbow Books, 2005. `http://www.ahds.ac.uk/guides/linguistic-corpora/index.htm`.

[27] E. Duval, W. Hodgins, S. Sutton, and S. L. Weibel, "Metadata principles and practicalities," *D-lib Magazine*, vol. 8, no. 4, p. 16, 2002. `http://www.dlib.org/dlib/april02/weibel/04weibel.html`.

[28] B. Hughes, "Metadata quality evaluation: Experience from the open language archives community," in *Digital Libraries: International Collaboration and Cross-Fertilization*, pp. 320–329, Springer, 2005.

[29] D. Broeder, M. Kemps-Snijders, D. Van Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, and C. Zinn, "A data category registry-and component-based metadata framework.," in *Proceedings of LREC 2010*, 2010. `http://www.windhouwer.nl/menzo/professional/papers/metaData.pdf`.

[30] Wikipedia, "Text Encoding Initiative (TEI)." `https://en.wikipedia.org/wiki/Text_Encoding_Initiative`, 2013.

[31] Steven Krauwer, "CLARIN ERIC: Common Language Resources and Technology Infrastructure." `http://www.clarin.eu/`, 2013.

[32] P. Wittenburg and D. van Uytvanck, "Chapter 2. Metadata: The Component Metadata Initiative (CMDI)." http://media.dwds.de/clarin/userguide/text/metadata_CMDI.xhtml, 2013.

[33] P. Boersma, "Scripting 6.4. Files." http://www.fon.hum.uva.nl/praat/manual/Scripting_6_4__Files.html, 2013.

[34] J. Meloni, "A gentle introduction to version control." http://chronicle.com/blogs/profhacker/a-gentle-introduction-to-version-control/23064, 2010.

[35] E. Carter, "The everyday developer's guide to version control with Git." http://www.slideshare.net/erincarter/the-everyday-developers-guide-to-version-control-with-git, 2009.

[36] Apache, "Apache HTTP Server Version 2.4 Documentation." http://httpd.apache.org/docs/2.4/, 2013.

[37] L. Corti, A. Day, and G. Backhouse, "Confidentiality and informed consent: Issues for consideration in the preservation of and provision of access to qualitative data archives," *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, vol. 1, no. 3, 2000. http://www.qualitative-research.net/index.php/fqs/article/view/1024.

[38] ICH Steering Committee, "Guideline for good clinical practice E6(R1)." https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6/E6_R1_Guideline.pdff, 2012.

[39] Wikipedia, "Open Science Data." https://en.wikipedia.org/wiki/Open_science_data, 2013.

[40] "Free Software Foundation." http://www.fsf.org/licensing/, 2013.

[41] "Open Source Initiative." http://opensource.org/, 2013.

[42] "Creative Commons." http://creativecommons.org/, 2013.

[43] European Union, "General Data Protection Regulation, GDPR." https://gdpr-info.eu/, 2016.

[44] Allen&Overy LLP, "The EU general data protection regulation." http://www.allenovery.com/SiteCollectionDocuments/Radical%20changes%20to%20European%20data%20protection%20legislation.pdf, 2017.

[45] ARTICLE 29 DATA PROTECTION WORKING PARTY, "Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is likely to result in a high risk for the purposes of Regulation 2016/679." https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/guidelines_on_data_protection_impact_assessment_dpia.pdf, 2017.

[46] IAPP, "The top 10 operational impacts of the EUs General Data Protection Regulation." https://iapp.org/resources/article/top-10-operational-impacts-of-the-gdpr, 2016.

[47] Information Commissioner's Office (ico.), UK, "Conducting privacy impact assessments code of practice." https://ico.org.uk/media/for-organisations/documents/1595/pia-code-of-practice.pdf, 2014.

[48] I. Wagner and E. Boiten, "Privacy risk assessment: From art to science, by metrics," *arXiv preprint arXiv:1709.03776*, 2017.

[49] D. Wright, R. Finn, and R. Rodrigues, "A comparative analysis of privacy impact assessment in six countries," *Journal of Contemporary European Research*, vol. 9, no. 1, 2013.

[50] A. Vocht, "The New EU General Data Protection Regulation and its Consequences for IT Operations and Governance." https://www.sqs.com/_resources/whitepaper-new-eu-general-data-protection-regulation.pdf, 2016.

[51] ico., "Overview of the general data protection regulation (gdpr)." https://ico.org.uk/for-organisations/data-protection-reform/overview-of-the-gdpr/, 2017.

[52] L. Sweeney, “Simple demographics often identify people uniquely,” *Health (San Francisco)*, vol. 671, pp. 1–34, 2000.

[53] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, “Unique in the crowd: The privacy bounds of human mobility,” *Scientific reports*, vol. 3, p. 1376, 2013.

[54] J. Henriksen-Bulmer and S. Jeary, “Re-identification attacksa systematic literature review,” *International Journal of Information Management*, vol. 36, no. 6, pp. 1184–1192, 2016.

[55] H. Ye, X. Cheng, M. Yuan, L. Xu, J. Gao, and C. Cheng, “A survey of security and privacy in big data,” in *Communications and Information Technologies (ISCIT), 2016 16th International Symposium on*, pp. 268–272, IEEE, 2016.

[56] U. D. of Health, H. Services, *et al.*, “Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (hipaa) privacy rule.” https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf, 2012.

[57] L. Presser, M. Hruskova, H. Rowbottom, and J. Kancir, “Care. data and access to uk health records: patient privacy and public trust,” *Technol. Sci.*, 2015.

[58] J. M. Rumbold and B. K. Pierscionek, “A critique of the regulation of data science in healthcare research in the european union,” *BMC medical ethics*, vol. 18, no. 1, p. 27, 2017.

[59] I. Budin-Ljøsne, P. Burton, J. Isaeva, A. Gaye, A. Turner, M. J. Murtagh, S. Wallace, V. Ferretti, and J. R. Harris, “DataSHIELD: an ethically robust solution to multiple-site individual-level data analysis,” *Public health genomics*, vol. 18, no. 2, pp. 87–96, doi:10.1159/000368959, 2015.

[60] DataSHIELD, “DataSHIELD – an open source solution for analysing sensitive data.” https://www.datashield.ac.uk, 2017.

[61] A. Gaye, Y. Marcon, J. Isaeva, P. LaFlamme, A. Turner, E. M. Jones, J. Minion, A. W. Boyd, C. J. Newby, M.-L. Nuotio, R. Wilson, O. Butters, B. Murtagh, I. Demir, D. Doiron, L. Giepmans, S. E. Wallace, I. Budin-Ljsne, C. Oliver Schmidt, P. Boffetta, M. Boniol, M. Bota, K. W. Carter, N. deKlerk, C. Dibben, R. W. Francis, T. Hiekkalinna, K. Hveem, K. Kvaly, S. Millar, I. J. Perry, A. Peters, C. M. Phillips, F. Popham, G. Raab, E. Reischl, N. Sheehan, M. Waldenberger, M. Perola, E. van den Heuvel, J. Macleod, B. M. Knoppers, R. P. Stolk, I. Fortier, J. R. Harris, B. H. Woffenbuttel, M. J. Murtagh, V. Ferretti, and P. R. Burton, “DataSHIELD: taking the analysis to the data, not the data to the analysis,” *International Journal of Epidemiology*, vol. 43, no. 6, pp. 1929–1944, doi:10.1093/ije/dyu188.

[62] R. Wilson, O. Butters, D. Avraam, J. Baker, J. Tedds, A. Turner, M. Murtagh, and P. Burton, “DataSHIELD–new directions and dimensions,” *Data Science Journal*, vol. 16, doi:10.5334/dsj-2017-021, 2017.

[63] S. M. Plis, A. D. Sarwate, D. Wood, C. Dieringer, D. Landis, C. Reed, S. R. Panta, J. A. Turner, J. M. Shoemaker, K. W. Carter, *et al.*, “COINSTAC: a privacy enabled model and prototype for leveraging and processing decentralized brain imaging data,” *Frontiers in neuroscience*, vol. 10, pp. Article 365, doi:10.3389/fnins.2016.00365, 2016.

[64] K. W. Carter, R. W. Francis, K. Carter, M. Bresnahan, M. Gissler, T. Grnborg, R. Gross, N. Gunnes, G. Hammond, M. Hornig, C. Hultman, J. Huttunen, A. Langridge, H. Leonard, S. Newman, E. Parner, G. Petersson, A. Reichenberg, S. Sandin, D. Schendel, L. Schalkwyk, A. Sourander, C. Steadman, C. Stoltenberg, A. Suominen, P. Surn, E. Susser, A. Sylvester Vethanayagam, and Z. Yusof, “ViPAR: a software platform for the virtual pooling and analysis of research data,” *International Journal of Epidemiology*, vol. 45, no. 2, pp. 408–416, doi:10.1093/ije/dyv193, 2016.

[65] Academy of Finland, “Detailed academy data management plan guidelines and best practices in dmptuuli.” http://www.aka.fi/en/funding/how-to-apply/application-guidelines/detailed-academy-data-management-plan-guidelines-and-best-practices-in-dmptuuli/, 2017.

[66] C. Dittrich, A. Negrouk, and P. G. a. Casali, "An ESMO-EORTC position paper on the EU clinical trials regulation and EMA's transparency policy: making european research more competitive again," *Annals of Oncology*, vol. 26, no. 5, pp. 829–832, doi:10.1093/annonc/mdv154, 2015.

[67] European Parliament and the Council of the EU, "Clinical trials regulation 536/2014." `https://ec.europa.eu/health/sites/health/files/files/eudralex/vol-1/reg_2014_536/reg_2014_536_en.pdf`, 2014.

[68] G. Chassang, "The impact of the eu general data protection regulation on scientific research," *ecancermedicalscience*, vol. 11, pp. 709, doi:10.3332/ecancer.2017.709, 2017.