



AUSSDA

AUSTRIAN  
SOCIAL SCIENCE  
DATA ARCHIVE


# DATA DEPOSIT GUIDELINE (Public version) v2.0

Information for Data Depositors

*"We make social science data accessible and reusable."*

30.06.2022

*Iris Butzlaff*

<b>Date</b>	30.06.2022
<b>Version</b>	2.0
<b>Licence</b>	 This work is licensed under a <a href="https://creativecommons.org/licenses/by/4.0/">Creative Commons Attribution 4.0 International Licence</a> .
<b>Access</b>	Public
<b>Suggested citation</b>	Butzlaff, Iris (2022). Data Deposit Guideline (Public version) v2.0. Vienna: The Austrian Social Science Data Archive.
<b>Contact</b>	University of Vienna Vienna University Library and Archive Services AUSSDA -The Austrian Social Science Data Archive Universitätsring 1 1010 Vienna Austria  T +43 1 4277 15323 <a href="mailto:info@aussda.at">info@aussda.at</a>

# Data Deposit Guideline

---

## Structure

Welcome – About AUSSDA.....	4
AUSSDA Dataverse.....	4
Workflow from data submission to publication .....	5
Available licence agreements and access options at AUSSDA .....	5
What you accomplish with a SUF licence agreement .....	6
How to prepare data for submission to AUSSDA .....	6
Gather and submit documentation material.....	6
Check the preferred formats.....	7
Check consistency and quality of data .....	8
Check the condition of the data files.....	9
Anonymisation/pseudonymisation of data for SUF licence .....	10
Technical requirements .....	12
Useful Stata commands for own data checks .....	12
Checklist for depositors .....	13

## Welcome – About AUSSDA

---

AUSSDA – The Austrian Social Science Data Archive is a data infrastructure for the social science community in Austria and offers a variety of research support services. We make social science data accessible, create opportunities for research and data re-use, benefitting science and society.

We operate as a professional archive for storing and/or disseminating your data for re-use as described in your data management plan. We are happy to accept data that are suitable either for research purposes (research data as well as replication data) or for teaching purposes.

We offer advice on the best sharing strategy for your data. We provide solutions for sensitive data that require restricted access as well as for data that should be accessible as openly as possible ("open access"). Together we will find a solution that meets your requirements and the open data or open access policies your funding agency or employer may have.

AUSSDA enhances the re-usability of your data: We assign DOIs (Digital Object Identifier) and help preparing the data and documentation material in a way that facilitates and promotes re-use. AUSSDA enriches your data with metadata so that the data can be found by interested researchers easily. In addition, AUSSDA is part of CESSDA ERIC, the Consortium of European Social Science Data Archives (European Research Infrastructure Consortium), the AUSSDA data catalogue is part of the CESSDA data catalogue. Therefore, the metadata related to your data will not only be visible (and searchable) in the AUSSDA data catalogue, but also on a European level in the CESSDA data catalogue and thus become part of the European Science Cloud (EOSC).

We use metadata standards for the description of data in social sciences (Data Documentation Initiative [DDI] and CESSDA controlled vocabularies) and store the information in the common data catalogue of CESSDA ERIC, so citing is facilitated and the data is easy to find. AUSSDA has been certified with the Core Trust Seal as a trusted repository since July 2020.

## AUSSDA Dataverse

The AUSSDA repository for research data and related material is called AUSSDA Dataverse and the software is based on The Dataverse Project - Dataverse.org. It is a web application to share, preserve, cite, explore and analyse research data. A collection of datasets around a special topic or source –also called a Dataverse – is a container for all your datasets, files, and metadata.

You can find all datasets in our [AUSSDA Dataverse](#).<sup>1</sup> The following list shows resources about using Dataverse:

AUSSDA Dataverse [User Guide](#)<sup>2</sup>

AUSSDA Dataverse [Datasets & files for download](#)<sup>3</sup>

---

<sup>1</sup> <https://data.aussda.at/>

<sup>2</sup> <https://aussda.at/en/aussda-dataverse-user-guide/>

<sup>3</sup> <https://aussda.at/en/rss/detail-en/news/dataverse-datasets-files-for-download/>



**Figure 1** Data depositing process (simplified presentation)

## Workflow from data submission to publication

AUSSDA offers advice how depositors can prepare their data for archiving in compliance with data protection (see also “Anonymisation/pseudonymisation of data<sup>4</sup>”). We undertake data checks, data processing (e.g. translating the data into different software formats) and give feedback to you as depositor. After the data curation, files will be converted into long-term and community-used formats and stored in the archive (Preservation). Eventually, the data is ready for publication in our [AUSSDA Dataverse](#).<sup>5</sup> Finally, the data can be re-used by other researchers. AUSSDA ensures access to your data by providing the necessary technical and legal infrastructure so that users can download the data themselves. Figure 1 illustrates the workflow from the acquisition and transmission of data, documentation material, metadata and legal documents by the depositor to the repository to make data and documentation material accessible.

## Available licence agreements and access options at AUSSDA

Since most of the studies archived with AUSSDA contain information about individuals and therefore comprise private and/or sensitive information, it is usually necessary to anonymise or pseudonymise the data for the publication of datasets. Anonymisation or pseudonymisation protect research subjects against identification, fulfil the requirements of data protection, and meet ethical research standards. Depending on the condition of the data – namely if it contains identifying variables, we publish the data as either Open Access (OA; also known as Public Use) or Scientific Use (SU).

For publishing the data Open Access, we apply Creative Commons licences.<sup>6</sup> These licenses provide a simple, standardized way to grant permission to share and use the work under copyright law. We mainly use the CC BY license. The CC BY aims at the maximal re-use of the data. The CC BY licensed data is meant to be used by teachers, students, journalists, policymakers, and the public but is sometimes not informative enough for researchers. Another argument for a Creative Commons licence is, that there is no identifying data or sociodemographic information in the datasets.

<sup>4</sup> AUSSDA uses the term “anonymisation/pseudonymisation” because the necessary steps to generate data that fulfil data protection rules depend on the condition of the different data sets.

<sup>5</sup> Find more information on the AUSSDA Dataverse here: <https://aussda.at/userguide/>.

<sup>6</sup> <https://creativecommons.org/about/clicenses/>

Datasets that are published under a *SUF licence agreement* and are accessible after the login to the AUSSDA Dataverse (what we call *restricted account-based access*) are of interest for research and teaching. Often, sociodemographic information about the research objects is included in the data. The restricted account-based access requires a login via federated account (single-sign-on) or via a registered account. Users sign up in the AUSSDA Dataverse and can then download the data and documentation files themselves after agreeing to the SUF terms of use.

The *SUF licence agreement with restricted controlled access* comprehends data that contain some kind of personal or other sensitive data or cover a specifically vulnerable group of research objects. The restricted controlled-based access also requires a login via federated account (single-sign-on) or via a registered account. Users sign up in the AUSSDA Dataverse and can then request access to restricted datasets. An AUSSDA team member is involved in the delivery of the data as the data users need to complete and sign a form and prove their legitimate scientific interest in the data. Then, an AUSSDA member verifies the scientific legitimation of the applicants and eventually grants access to the data. After approval and while being logged in, the users can then download the data.

All licence agreements or types of access have in common that all *direct* identifiers (see examples below) have to be removed from the data before publication. This requirement is based on the EU Data Protection Regulation (GDPR)<sup>7</sup> and the Austrian Data Protection Act.<sup>8</sup> An exception for the scientific re-use of data including direct identifiers is possible if these direct identifiers have previously been lawfully published and the data stem from publicly available sources.

#### What you accomplish with a SUF licence agreement

The aim of the SUF (scientific use file) licence agreement is the re-use of a dataset by researchers who have a well-defined scientific research purpose.

## How to prepare data for submission to AUSSDA

### What AUSSDA expects from depositors:

The depositor assembles the archive material and prepares all documents and data according to our guidelines. This includes the collection of all documentation material (see list below), the comparison of documentation material with data (with regard to labelling, spelling, etc.), anonymisation/pseudonymisation procedures (see Table 2) and a check on technical requirements (see Table 3).

- We organize and rename all data and documentation files according to our internal file naming scheme.
- We add two additional standard variables to the dataset ("AUSSDA version number" and "DOI").

### Gather and submit documentation material

The submission of comprehensive documentation along with the data is crucial for the re-usability of data. The more documentation material is made available, the higher is the

---

<sup>7</sup> In Austria: Datenschutzgrundverordnung, DSGVO, according to European Law.

<sup>8</sup> In Austria: Österreichisches Datenschutzgesetz, DSG.

understanding for the re-using scientists. The following list provides an overview which files we consider as mandatory, as recommended and as optional documentation material.

- I. The following documents are mandatory for publication:
  - Licence agreement (signed by depositor and AUSSDA)
  - Data files
  - Instruments of data collection (e.g. questionnaire with interviewer instructions, information material for respondents, data collection guideline)
  - Metadata sheet (we need the metadata of your dataset to fill in the relevant information to make your data Findable, Accessible, Interoperable and Re-usable [FAIR])
  
- II. The following documents are strongly recommended:
  - Codebook
  - Method report
  
- III. The following documents are optional:
  - Project report
  - Data Management Plan (DMP) of project proposal
  - Interviewer guideline
  - Interview cards
  - Documentation about incentives, contacts
  - Recoding protocol
  - Informed consent forms
  - Any further document that helps users to understand the data

**Please make sure that you send two versions of your data:** the *dataset as you used it for your analysis* and the *anonymised/pseudonymised dataset* according to our anonymisation/pseudonymisation recommendations (Table 2). With the first mentioned dataset, we might have the opportunity to offer more detailed versions of your dataset in case legal requirements (e.g. data protection regulations) change or we can deliver single (sensitive) variables in a separate access procedure when necessary.

All files should be transferred to AUSSDA in a secure way (e.g. by the [Aconet Filesender](#)<sup>9</sup>). As we process all data and related documents (e.g. for data checks and the migration into long-term archive formats), all write protection must be removed before submission.

#### Check the preferred formats

In Table 1, we provide information about which data formats we prefer for submission to AUSSDA. This guideline ensures that we can preserve your data and guarantee accessibility for re-use in the future. If you have any questions concerning conversion or suitability of your data, do not hesitate to contact us! We update this list on a regular basis to account for software changes or disciplinary-specific format changes.

---

<sup>9</sup> <https://filesender.aco.net>.

**Table 1.** List of preferred formats

Data type	Preferred formats	Acceptable formats
Quantitative data	<ul style="list-style-type: none"> <li>○ Proprietary formats of statistical software, such as Stata 14 (.dta), SPSS (.sav)</li> <li>○ Tab-, or comma-delimited text files (e.g., .csv, .tab, .tsv, etc.) with command file (setup/syntax for import into Stata or SPSS)</li> <li>○ (only characters not in the data should be used as delimiters or appropriate encapsulation is needed)</li> </ul>	<ul style="list-style-type: none"> <li>○ OpenDocument table format (.ods), MS Excel (.xlsx, .xls)</li> <li>○ Tab-, or comma-delimited text files (e.g., .csv, .tab, .tsv) without command file</li> <li>○ Binary format</li> <li>○ Statistical software R (.R)</li> </ul>
Qualitative data, documentation, and programming scripts	<ul style="list-style-type: none"> <li>○ PDF/A (.pdf)</li> <li>○ Plain text, ASCII, UTF8 (.txt)</li> </ul>	<ul style="list-style-type: none"> <li>○ OpenDocument Text (.odt), MS Word (.docx, .doc)</li> <li>○ PDF (.pdf)</li> <li>○ Rich Text Format (.rtf)</li> <li>○ Markdown (.md)</li> <li>○ Hypertext Markup Language (.htm, .html)</li> <li>○ Extensible Markup Language (.xml)</li> <li>○ JavaScript Object Notation (.json)</li> </ul>
Images	<ul style="list-style-type: none"> <li>○ TIFF (.tif, .tiff) (but do not convert JPEG to TIFF, as this would not enhance the quality of the images)</li> </ul>	<ul style="list-style-type: none"> <li>○ JPEG (.jpeg, .jpg)</li> <li>○ PDF/A, PDF (.pdf)</li> <li>○ PNG (.png)</li> <li>○ SVG (.svg)</li> <li>○ BMP (.bmp)</li> </ul>

## Check consistency and quality of data

We highly recommend depositors to ensure that the following aspects apply to their data and their documentation material:

- The dataset complies to the documentation material (e.g. questionnaire/survey) and vice versa. Frequent errors are e.g. a lack of completeness and comprehensibility of labels.
  - Examples:
    - Variable f1\_4 has the options "don't know" and „not applicable" in the dataset, but the codebook does not contain these options.
    - The variable capturing the respondents' age is named "age\_respondent" in the dataset, but "Age" in the codebook. This discrepancy requires either renaming the variable or adapting the codebook.
- All outliers are valid.
- The observations are plausible.



- Example: If a respondent is 15 or younger, he/she could not have cast a vote in the last general election. If a respondent indicates to be single, the observation for the spouses' occupation should be either missing or not applicable.

## Check the condition of the data files

In order to fulfil data protection guidelines, we recommend steps to ensure that individuals cannot be re-identified.

*Direct identifiers* allow the re-identification of individuals very easily so their deletion is mandatory: Not only names or telephone numbers render individuals identifiable but also online identifiers such as IP addresses. An exception is possible if direct identifiers have been lawfully published previously and the data stem from publicly available sources.

With regard to *indirect identifiers*, most probably the combination of "one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person" (GDPR) can identify an individual, too. So, depending on the dataset, data protection may also require the recoding of demographic variables that contain outliers or low numbers of observations that facilitate re-identification in combination with other variables. Examples for such variables are detailed information of occupations (e.g. there are fewer legislators and generals than medical doctors or shop keepers) or respondents' origins.

*Open questions* involve the risk that respondents disclose details of their life and/or identity which allows the re-identification of the respondents. Therefore, checks if any details are disclosed in these variables are a precondition for submission and publication of the data. In case (text) variables contain personal details, this information coercively needs to be deleted or recoded. In case, text variables remain in the data set, the depositor needs to confirm to AUSSDA staff, that all variables have been checked individually for information that might lead to re-identification of respondents. The AUSSDA team undertakes random checks whether open answers have been checked and recoded or deleted. AUSSDA staff may also recommend a more restrictive access mode in this case if deemed sensible.

In case that the respondent group is especially vulnerable (like children, refugees, or ethnic and migrant minorities), we recommend strict adherence to the proposed pseudonymization steps.

Legal restrictions when depositing data at AUSSDA might emerge from laws and directives concerning intellectual property rights. We recommend to check our "[Guideline on Intellectual Property Rights](#)".<sup>10</sup>

---

<sup>10</sup> <https://aussda.at/aussda-ipr-guideline>

## Anonymisation/pseudonymisation of data for SUF licence

**Table 2:** Anonymisation/pseudonymisation procedure

Removal of all (in)direct identifiers	Example
Delete variable	Social security number ID number from 3rd party (data collection institute) Full name Email address Phone numbers Postal codes Date of birth as (DDMMYYYY) <sup>11</sup> Workplace/employer Vehicle registration number Bank account number IP address Student ID number Passport/identity card number
Answers to open questions	Example
Thoroughly check or delete answers to open questions	Do you have any experience in working in politics? <u>Answer:</u> I have been elected as delegate in the parliament for party XYZ for 12 years in a row. -> Delete or recode!
	Do you carry out any honorary duties? <u>Answer:</u> I have been working as a volunteer in the fire department in village XYZ in district ABC. -> Delete or recode!
Standard demographic variables	Example
Age -> summarize if margins fizzle out	Younger than 18 years 19 20 ... 70 and older
Occupational status -> categorize into groups with min. 20 observations	Employed Self-employed Student/in training/in school Retired Unemployed Other (e.g. military service, social service)
Field of education -> categorize into groups with min. 20 observations	Building and civil engineering Food processing

<sup>11</sup> Date of birth as MMYYY is possible to be delivered by controlled access procedure as additional variable.

Years of schooling -> summarize into groups with min. 20 observations and if margins fizzle out	Less than 5 years 6 7 8 ... 13 and more
Income -> categorize into broader categories and if margins fizzle out	Below 500 EUR 500 to below 1000 EUR 1000 to below 1500 EUR, etc. More than 4000 EUR (e.g.)
Affiliation to religious groups -> categorize into groups with min. 20 observations	Christian Jewish Islamic etc.
Membership in associations/clubs/political parties/trade union -> categorize into groups with min. 20 observations	Recode "member of dachshund breeders association (in Linz)" to a broader category
Household composition -> summarize into groups with min. 20 observations and if margins fizzle out	1 person 2 persons, etc.  > 5 persons
Nationality -> categorize into groups with min. 20 observations or make broader categories	Austrian  Ethiopian
Mother tongue -> categorize into groups with min. 20 observations	German  Kiswaheli
Need for social welfare -> categorize into groups with min. 20 observations	Program A  Program B
Drug abuse -> categorize into groups with min. 20 observations	Drug addiction: yes, no
Legal background information -> delete if less than 20 observations in one category	Judicially condemned: yes, no
Health information -> delete if less than 20 observations in one category	Suffer from depression: yes, no

Country of origin -> if less than 20 observations in one category, use standard for area codes used by the UN, UN M49 (if possible); use subregional category	Use UN geoscheme, e.g.  Eastern Africa, Middle Africa, Northern Africa, Southern Africa, Western Africa, etc.
ISCO variables on 3 <sup>rd</sup> level	3 digits level
NUTS variables on 2 <sup>nd</sup> level	Provinces, Bundesländer

## Technical requirements

AUSSDA offers data files in different formats. Stata files are stored in version 14 and in csv (comma-separated values) format because this ensures the long-term preservation of the data. Furthermore, AUSSDA offers data in SPSS format. All formats are uploaded in the AUSSDA Dataverse in unicode format. Therefore, submitted data has to comply to several technical requirements in order to avoid problems when AUSSDA staff converts the data to other formats.

**Table 3:** Technical requirements

Technical setting	Reason
system missing  should be coded numerical (e.g. by using negative values)  (e.g. -9 no answer; -8 do not know)	In Stata, system missing can be implemented as .a or .b etc. After conversion of the dataset to SPSS or R, these system missings are not displayed correctly any more.
use EUR instead of €	The € sign might cause problems when saving the data in unicode format.
variable labels should not be longer than 79 characters	Variable labels are truncated after 79 characters in Stata and remain incomplete. They may be not understandable.

## Useful Stata commands for own data checks<sup>12</sup>

- `notes` // check if there are notes attached to the datafile and delete these notes
- `codebook` // list variable names and values, range, missings, value labels, unique values and check if these correspond to the codebook
- `duplicates report id` // check if each observation has one unique id
- `isid id` // check if id is missing

<sup>12</sup> See also this do-file for a more detailed list of commands and explanations:

[https://www.aussda.at/fileadmin/user\\_upload/p\\_aussda/Documents/recommended\\_datachecks.do](https://www.aussda.at/fileadmin/user_upload/p_aussda/Documents/recommended_datachecks.do)

- `label list` // special characters in value labels (e.g. ß or €) may cause problems during conversions even if these value labels are not assigned to variables. We therefore recommend to delete all labels that are not used in the dataset.
- `labelbook [labelname]` // find out if a label is (not) assigned to a variable
- look for unlabelled values // labelling of variables facilitates re-use:
 

```
scandata, nolabel13
local varnolabels = r(mis_lab)
foreach varnolab of local varnolabels {
codebook `varnolab', tab(100)
}d
```
- `findname, type(string) local(strvars)` // find string variables, can then check these on sensitive/identifying information

### Checklist for depositors

- assemble all archive materials
- sign deposit contract
- compare documentation materials with data
- check the preferred formats
- check on spelling errors, correct labels
- undertake a pseudo-/anonymisation procedure
- fulfil technical requirements
- remove restrictions from all documents
- send archive material to AUSSDA

#### Archive materials include:

- signed licence agreement (contract)
- non-anonymised/pseudonymised dataset
- anonymised/pseudonymised dataset
- metadata sheet
- questionnaire(s) (in different languages\*)
- codebook\*
- method report (incl. undertaken anonymisation/pseudonymisation measures)\*
- confirmation of undertaken curation of sensitive text variables (string variables)\*

\* = optional

Please do not hesitate to contact us under [info@aussda.at](mailto:info@aussda.at), if you have questions or requirements not covered in this document – we are happy to assist you in finding a solution.

<sup>13</sup> It is necessary to install this command before using: `ssc install scandata`.