AUSSDA

AUSTRIAN
SOCIAL SCIENCE
DATA ARCHIVE

PRESERVATION PLAN

“We make social science data accessible and reusable.”

01.12.2022

Christian Bischof

Date	01.12.2022
Version	1.0
License	This work is licensed under a Creative Commons Attribution 4.0 International License.
Access	Public
Suggested citation	Bischof, Christian (2022). Preservation Plan. V1.0. Vienna: The Austrian Social Science Data Archive.
Contact	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 info@aussda.at

Preservation Plan

This document describes the preservation plan of AUSSDA - The Austrian Social Science Data Archive. It addresses the archive's approach to preserving data deposited by members of its users. In depth information on acceptable data (formats) and the designated user community can be found in the AUSSDA Data Collection Policy. This preservation plan is structured along the process of data preservation, from the ingest stage to the stage of providing archive users with access to the data. According to the reference model for an Open Archival Information System (OAIS) the preservation planning requires the monitoring of the designated community.

AUSSDA monitors the needs of the designated community, members of all disciplines of the social sciences, with the following measure

- The In-house Governing Board ("Leitungsgremium") directly supports the management of AUSSDA. It advises on questions of strategy and future developments. The board consists of representatives of the Universities of Vienna, Graz, Linz and Innsbruck, and the Austrian Federal Ministry of Education, Science and Research.
- The External User Advisory Board provides a communication channel to our designated community to determine the future direction of the archive and its services.
- Ongoing contact during everyday operation with data depositors and data re-user.

From these sources AUSSDA gets information about new requirements concerning data and file formats, software preferences etc. From these requirements AUSSDA is developing new preservation strategies for long-term preservation.

The general approach is to provide access to the data holdings in widely used formats. Recommended file formats[1] are defined, for these formats the archive ensures preservation of the data to make it accessible for reuse in the future. AUSSDA uses the two most common statistical software formats in social sciences, Stata and SPSS. These data formats are the output of the preservation process; tab-separated values (TSV) ASCII (UTF-8) text file as additional data format is also available because it can be used with non-proprietary software. Documentation material is stored as PDF/A (a suitable format for long-term preservation of page-oriented documents) or as ASCII (UTF-8) text file.

Preservation process

The different stages of the preservation process follow the OAIS reference model. The deposited data is processed with the concept of information packages:

- Submission Information Package (SIP): An information package that is delivered by the data producer to the archive for use in the construction or update of an AIP.
- Archival Information Package (AIP): An information package, consisting of the content information and the associated preservation description information.
- Dissemination Information Package (DIP): An information package, derived from AIP, and transferred to the user in response to a request.

File naming and file organization schemas are described in an internal document. All filenames of processed files in AIP and DIP have to follow a file naming scheme, which defines a unique archive

[1] https://aussda.at/en/faq-downloads/faq/#c798251 -> Recommend Formats

number, the type of file (data, questionnaire, method report etc.), the language of the file, and the version (two levels: major/ minor changes).

In order to comply with the requirements of a documented preservation process of data and documentation files, we have established a Data Lifecycle Log, where the locations of all files with their description are stated. The process of preservation (data transformation processes, checks, and final storage) is also described in the DLC Log. Additional changes are documented and versioned.

Workflow

Figure 1 shows the workflow of the preservation process[2]. All data and documentation material from the data depositor is stored in the SIP, this package is the source of any further processing and will never be modified. In the processing stage, the packages are on the working volumes.

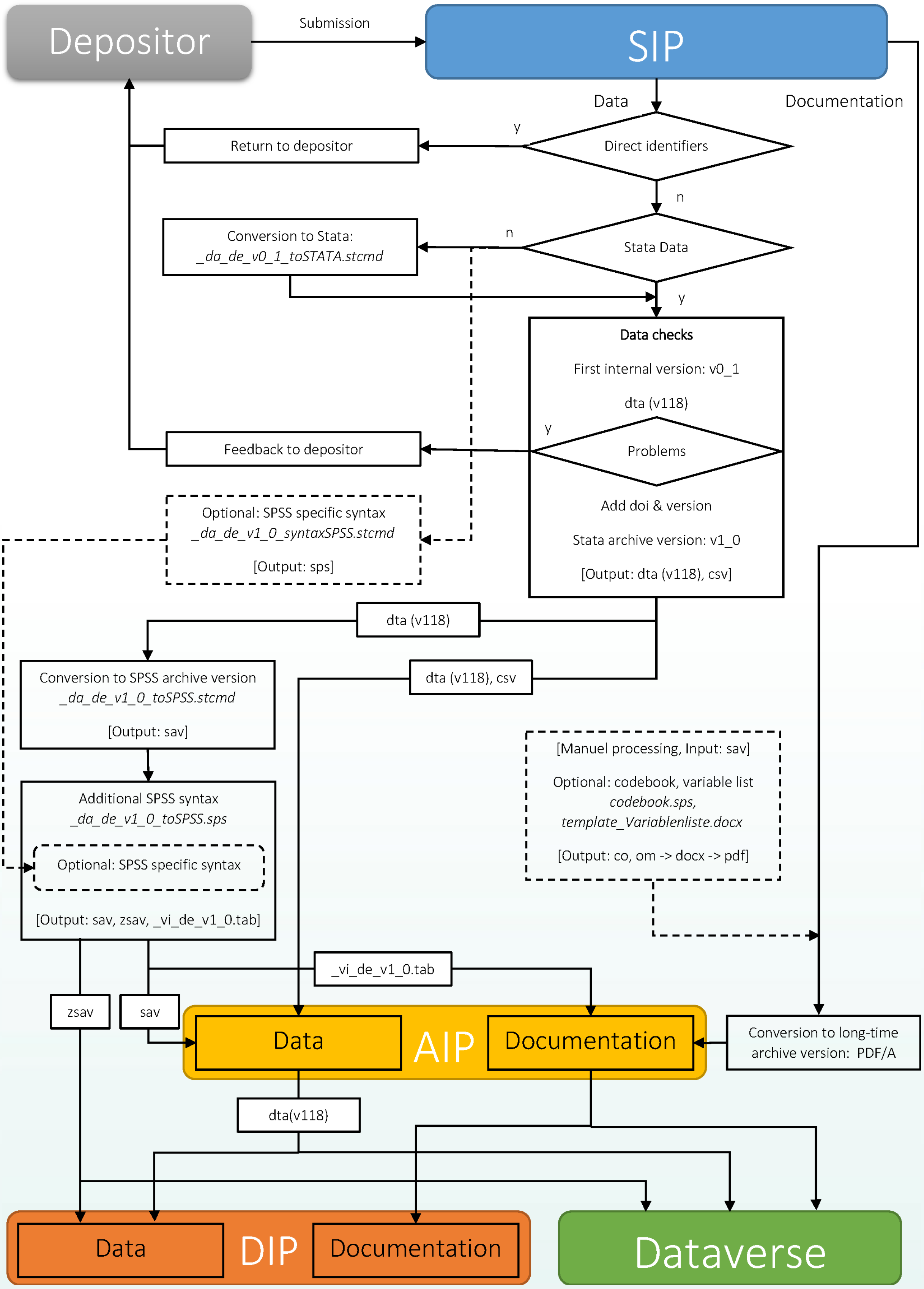
The first step is to check whether the submitted material contains direct identifiers: full name, email address, phone number, postal code, data of birth, social security number etc. This information could be used to identify, locate, or reveal the characteristics of other details about individuals. If any of this information is included in the submitted material, AUSSDA deletes the material and requests an update from the depositor. The depositor has to clean the material and submit it again.

The second step is to check the format of the data. If the data is not in the standard processing format Stata (dta), the data will be converted with the software Stat/Transfer[3] into the Stata format. The Stat/Transfer command file (stcmd) is stored like all other processing files in the AIP package folder.

The main ingest processing is done with Stata, all commands are stored in a Stata command file (do-file). The processing can be reproduced, changes can be easily made and the processing files can be re-run. The main purpose of the ingest is the comparison of data and documentation material (labels, values, etc.), plausibility checks and data protection checks. For an easy implementation of all tasks a data check template Stata do-file is in use. The do-file template follows the data preparation recommendation for the data depositors, they are part of the Data Deposit Guideline[4]. If problems or questions occur, they are reported back to the data depositor in order to find a solution together and to process the data according to the archive standards as well as to guarantee understandability to re-users of the data. As one of the last steps in the main ingest process two variables are added to the dataset, the number of the dataset as digital object identifier (DOI) and the version of the dataset (including date), this uniquely identifies the dataset. The dataset is stored as Stata dta (specification 118) file and as comma-separated value ASCII text file in the AIP folder.

[2] All steps are detailed documented through the only internally accessible AUSSDA Wiki.
[3] Stat/Transfer software https://stattransfer.com
[4] AUSSDA Data Deposit Guideline
https://www.aussda.at/fileadmin/user_upload/p_aussda/Documents/Data-Deposit-Guideline_SUF_v2_0.pdf

Figure 1


Depositor
Submission
SIP
Data
Documentation
y
Return to depositor
Direct identifiers
n
Conversion to Stata: _da_de_v0_1_toSTATA.stcmd
n
Stata Data
y
Data checks
First internal version: v0_1
dta (v118)
y
Feedback to depositor
Problems
Add doi & version
Stata archive version: v1_0
[Output: dta (v118), csv]
Optional: SPSS specific syntax _da_de_v1_0_syntaxSPSS.stcmd
[Output: sps]
dta (v118)
Conversion to SPSS archive version _da_de_v1_0_toSPSS.stcmd
[Output: sav]
dta (v118), csv
[Manuel processing, Input: sav]
Optional: codebook, variable list codebook.sps, template_Variablenliste.docx
[Output: co, om -> docx -> pdf]
Additional SPSS syntax _da_de_v1_0_toSPSS.sps
Optional: SPSS specific syntax
[Output: sav, zsav, _vi_de_v1_0.tab]
_vi_de_v1_0.tab
zsav
sav
Data
AIP
Documentation
Conversion to long-time archive version: PDF/A
dta(v118)
Data
DIP
Documentation
Dataverse

For a data format conversion from Stata to SPSS again Stat/Transfer is used, the data file (sav) and the processing file (stmcd) are also saved in the AIP folder. A conversion from SPSS to Stata looses some SPSS specific data format properties. Because of different data format specifications between SPSS and Stata in a next step, we optionally recover some SPSS properties[5]. The SPSS specific instruction is generated with Stat/Transfer and inserted in a SPSS syntax file (sps). The SPSS syntax files stores the data files in the AIP (sav) and DIP (compressed SPSS zsav) folders. Additionally, a documentation file with machine-readable variable identifiers and descriptions is generated and stored. After processing of all data files, the SPSS (zsav) and Stata (dta) data files are uploaded to the Dataverse repository system by AUSSDA staff. Dataverse performs an ingest process [6], the files are processed and converted into the archival format of the Dataverse application. The Dataverse installation stores the raw data content extracted from the proprietary data formats in a plain text, tab-delimited format. To prevent the ingest from the SPSS file, the compressed zsav file is used, this format will not be ingested by Dataverse[7]. The metadata information that describes the file content is stored separately, in a relational database, so that it can be accessed efficiently by the Dataverse application. The data file metadata describes the ingested data vectors, it represents the observation values of a variable with descriptive variable and value label. Dataverse also records descriptive metadata about datasets (a collection of data and documentation files). AUSSDA uses the CESSDA Metadata Model[8], this is derived from the social science metadata standard of the Data Documentation Initiative (DDI) 2.5[9].

The documentation material is converted to the PDF/A format, the ISO standard for long-term archiving [10] and stored in the AIP and DIP documentation folders. The content of the DIP documentation folder is uploaded to the Dataverse repository system and makes the data holdings accessible for users.

After all processing tasks and the successful publication of the dataset have been completed, finally the complete data packages (SIP, AIP, DIP) are moved to the functional entity of the archival storage volumes, and access rights for non-preservation employees get restricted to read-only. The archival storage is maintained according to a backup and recovery plan. If a revision of the data is necessary, the data packages will be transferred to the working volumes. After the revision is done, the packages are transferred back on to the archival storage volumes again.

Used data formats

After processing the research data is stored in a variety of data formats. A user has access to formats that are made available with Dataverse. These files are also stored in the DIP. The AIP holds the versions for long-term preservation.

	AIP	DIP	Dataverse
Stata (.dta, specification 118)	x	x	x
SPSS (.sav, UTF-8 encoded)	x		
SPSS (.zsav, UTF-8 encoded)		x	x
comma-separated value (.csv)	x		
tab-separated values (.tsv)			x

[5] Numeric missing codes, variable value labels longer than 80 characters.
[6] https://guides.dataverse.org/en/latest/user/tabulardataingest/index.html
[7] SPSS does not openly publish the specifications of their proprietary file formats. Because of that Dataverse cannot fully guarantee the ingest.
[8] https://zenodo.org/record/4751455#.Y2tsSuSZNaR
[9] https://ddialliance.org/Specification/DDI-Codebook/2.5
[10] https://www.iso.org/standard/71832.html

Digital Preservation

Digital preservation ensures that digital information remains accessible and usable. AUSSDA does not specify a maximum storage period, i.e. that the data will be archived "forever". In principle, the AUSSDA transfer and license agreement[11] does not provide for the deletion of data. The only reasons for deletion are of a legal nature. Through the use of a persistent identifier the findability of the published data and documentation material is guaranteed. AUSSDA uses the DOI (digital object identifier) at study level. The DOI ensures that the links to the study metadata are resolved, even if the data have to be de-accessioned because of legal reasons.

Digital preservation is implemented on three levels at AUSSDA:

Level 1: Bit-level preservation is the lowest level of preservation; the preservation of the binary information 0 and 1 on the physical layer. The data holdings are duplicated and the copies are physically stored separately. The storage devices are maintained and regularly replaced. Fixity checks are in place to guarantee the bit-level integrity. The bit-level preservation is within the responsibility of the Vienna University Computer Center (ZID).

Level 2: Logical preservation ensures that digital data is preserved in a technically processable and readable manner, regardless of any technological changes. The AUSSDA approach is to migrate to new formats, when formats currently in use become obsolete[12]. At the moment, AUSSDA supports the two most common data formats in the social sciences.The latest specification of Stata (dta, specification 118)[13] and SPSS (sav, unicode)[14] are in use. Both formats can be imported from other applications. Additionally a format with tab-separated values (plus descriptive variable and value label information) is available. Documentation material generally is available in the long-term archiving format PDF/A. When new formats become necessary due to technical obsolescence or new requirements of the designated community, AUSSDA is able to perform format conversions. AUSSDA has different possibilities to create new data formats:

- New format specification for Stata or SPSS through the native application.
- Conversion to additional formats with Stat/Transfer.
- Ingested archival format of Dataverse application stores separated the raw data as tab-separated values file on the one hand and descriptive data about variables and values on the other hand. This information can be used by new tools to generate new data formats.[15]
- The AIP holds raw data as comma-separated value file as well as variable identifiers and descriptions separately. Through this information new data formats can be generated.

Because of the consistent and structured file naming and file organization structure of all data holdings, necessary conversion operations can be carried out automatically in the future[16].

[11] AUSSDA Transfer and license agreement
https://aussda.at/fileadmin/user_upload/p_aussda/Documents/AUSSDA_DepositAgreement_su_v1_4_en.pdf
[12] The emulation of software or hardware is not necessary because the AUSSDA data formats can be used with different application on different hardware platforms.
[13] https://www.loc.gov/preservation/digital/formats/fdd/fdd000471.shtml
[14] https://www.loc.gov/preservation/digital/formats/fdd/fdd000469.shtml
[15] https://guides.dataverse.org/en/latest/api/external-tools.html
[16] Until now, it has not been necessary to perform such large-scale operations. Nevertheless, data preparation makes it sometimes necessary to perform batch processing e.g. conversion from older Stata specification to specification 118.

Level 3: Semantic preservation ensures understandability and interpretability of the data to re-users. Semantic preservation is fulfilled through descriptive metadata and archived documentation material. The study documentation explains the aims and the context of the data as well as the research questions, used methodologies, how data was collected, the used instruments and measures, etc.

Validity of the Preservation Plan

The preservation plan has been reviewed and approved by staff working in the areas of ingest, preservation and access and by the head of AUSSDA.