DATA INGEST PROCESSING STANDARDS

PUBLIC
17 NOVEMBER 2014
Version: 07.00

T +44 (0)1206 872001
E sharonb@essex.ac.uk
www.data-archive.ac.uk

UK DATA ARCHIVE
UNIVERSITY OF ESSEX
WIVENHOE PARK
COLCHESTER
ESSEX, CO4 3SQ

WE ARE SUPPORTED BY THE **UNIVERSITY OF ESSEX**, THE **ECONOMIC AND SOCIAL RESEARCH COUNCIL**, AND THE **JOINT INFORMATION SYSTEMS COMMITTEE**

Contents

Scope

What is included in this document

This document defines data collection ingest processing standards currently in use at the UK Data Archive, and the criteria used to allocate those standards.

What is not included in this document

This document does not include details of quantitative, qualitative or Secure Access data ingest processing procedures. Further details of procedures may be found in the documents *Quantitative Data Ingest Processing Procedures, Qualitative Data Ingest Processing Procedures and Secure Access Ingest Procedures*.

It should be noted that some of the documents referenced within the text below are not publicly available, but external readers may of course contact the Archive in case of query.

1. The ingest processing standard and the data access level

The ingest processing standard allocated is based solely on the condition of the data and documentation and the anticipated level of secondary use (see section 2 below). It is not affected by and should not be confused with the *access level* of the data collection (End User Licence, Special Licence or Secure Access).

2. Allocation of the ingest processing standard

All data collections added to the Archive collection are ingest processed to one of four set standards (A*, A, B or C), which determine the level of work to be undertaken to prepare them for secondary use. All quantitative and qualitative data collections undergo confidentiality checks irrespective of the processing standard allocated. The confidentiality checks undertaken are commensurate with the access level of the data collection (see section 1 above) regardless of the ingest processing standard.

For both quantitative studies and qualitative data collections, allocation of an ingest processing standard depends largely on:

- condition of the data and documentation;

- anticipated level of secondary use.

The ingest processing standard is therefore usually allocated based on the results of pre-processing checks carried out on the study component files.

However, some set rules for processing standard allocation do exist:

- Quantitative studies destined for the Archive's Nesstar browsing tool - will be processed to A* standard (see list in Appendix).
- Quantitative studies deposited by the Office for National Statistics (ONS) or another government department are generally processed to A standard if in suitable condition, with a few exceptions.
- Quantitative studies deposited by a major research centre such as NatCen Social Research (NatCen) are generally processed to A standard if in suitable condition.
- Secure Access studies are usually processed to either A or B standard depending on the condition of the data and documentation at time of deposit. At present, all Secure Access studies in the collection are quantitative in nature.
- If the quantitative study is part of a series (regular or otherwise), the processing standard allocated should be that previously used for the series. It is desirable to use the same standard in order to keep the series consistent, but if the condition of the study components received have improved (or markedly deteriorated) since the last deposit, the standard may be raised (or lowered).
- Most quantitative studies originating from academic sources (where added to the main Archive collection rather than ESRC-store) will be processed to B standard or higher. However, if the study components are in suitably good condition, or a high level of reuse is anticipated, a higher processing standard may be allocated.
- Quantitative processing standard C is only used for those studies where the materials are in very poor condition with little improvement possible, or are in a software-dependent format with no alternatives available (the depositor should be contacted at the pre-processing check stage (or at Pre-Ingest) to ascertain whether extra or better materials are available).
- Qualitative and mixed-methodology data collections destined for online enhancements (such as the Qualibank system) will be processed to qualitative processing standard A*. Otherwise, they may be processed to qualitative A or B standard depending on condition and anticipated re-use.

3. Quantitative ingest processing standards

The following sections describe the quantitative ingest processing work undertaken for each processing standard (A* to C).

3.1. Processing standard A*

The text below describes the quantitative ingest processing work undertaken for processing standard A*. This text (or a variant thereof) appears in the 'Read' file associated with each quantitative study processed to A* standard:

"The data were processed to the UK Data Archive's 'A*' standard. This is the Archive's highest standard, and means that an extremely rigorous and comprehensive series of checks was carried out to ensure the quality of the data and documentation. Briefly, the most important procedures were as follows. Firstly, checks were made that the number of cases and variables matched the depositor's records. Secondly, checks were made that all variables had comprehensible variable labels and all nominal (categorical) variables had comprehensible value labels. Where possible, either with reference to the documentation and/or in communication with the depositor, labels were accordingly edited or created. Thirdly, logical checks were performed to ensure that nominal (categorical) variables had values within the range defined (either by value labels or in the depositor's documentation). Lastly, any data or documentation that breached confidentiality rules were altered or suppressed to preserve anonymity.

All notable and/or outstanding problems discovered are detailed under the ‘Data and documentation problems’ heading below.”

3.2. Processing standard A

The text below describes the quantitative ingest processing work undertaken for processing standard A. This text (or a variant thereof) appears in the ‘Read’ file associated with each quantitative study processed to A* standard:

“The data were processed to the UK Data Archive's ‘A’ standard. A rigorous and comprehensive series of checks was carried out to ensure the quality of the data and documentation. The most important procedures were as follows. Firstly, checks were made that the number of cases and variables matched the depositor's records. Secondly, checks were made that all variables had variable labels and all nominal (categorical) variables had value labels. Where possible, either with reference to the documentation and/or in communication with the depositor, absent labels were created. Thirdly, logical checks were performed to ensure that nominal (categorical) variables had values within the range defined (either by value labels or in the depositor's documentation). Lastly, any data or documentation that breached confidentiality rules were altered or suppressed to preserve anonymity.

All notable and/or outstanding problems discovered are detailed under the ‘Data and documentation problems’ heading below.”

3.3. Processing standard B

The text below describes the quantitative ingest processing work undertaken for processing standard B. This text (or a variant thereof) appears in the ‘Read’ file associated with each quantitative study processed to B standard:

“The data were processed to the UK Data Archive's B standard. A substantial series of checks was carried out to ensure the quality of the data and documentation. Firstly, checks were made that the number of cases and variables matched the depositor's records. Secondly, logical checks were performed on a sample of the remaining nominal (categorical) variables to ensure they had values within the range defined (either by value labels or in the depositor's documentation). Thirdly, any data or documentation that breached confidentiality rules were altered or suppressed to preserve anonymity.

All notable and/or outstanding problems discovered are detailed under the ‘Data and documentation problems’ heading below.”

Note: checks performed at B standard are dependent on the nature of the study.

3.4. Processing standard C

The text below describes the quantitative ingest processing work undertaken for processing standard C. This text (or a variant thereof) appears in the ‘Read’ file associated with each quantitative study processed to C standard:

“The data were processed to the UK Data Archive's ‘C’ standard, which is the Archive's minimal processing level. A series of checks was carried out to ensure the quality of the data. Checks were made that the number of cases and variables matched the depositor's records. Any data or documentation that breached confidentiality rules were altered or suppressed to preserve anonymity.

All notable and/or outstanding problems discovered are detailed under the ‘Data and documentation problems’ heading below.”

4. Qualitative ingest processing standards

The following sections illustrate typical ingest processing work undertaken on qualitative data collections. Further details of qualitative ingest processing procedures may be found in the document *Qualitative Data Ingest Processing Procedures*.

4.1. Processing standard A*

- If not born digital, data are fully digitised and anonymised.
- Data are accessible online and via download from the UK Data Archive
- Some or all of the data may never have been available in digital format before
- Metadata are fully digitised and anonymised.
- Enhanced metadata are accessible online and via the Archive's catalogue.
- Additional work may have been done to create new metadata that adds to understanding of the collection

4.2. Processing standard A

- If not born digital, data are fully digitised and anonymised
- Data are accessible through UK Data Archive
- Metadata are fully digitised and anonymised
- Metadata are accessible through UK Data Archive

4.3. Processing standard B

- If not born digital, data are digitised (at least to the level of scanned TIFF images) and anonymised
- Only major problems with data are resolved
- Data are accessible through UK Data Archive
- If not born digital, metadata are digitised at least to the level of scanned TIFF images and anonymised
- Only major problems with metadata are resolved
- Metadata are accessible through UK Data Archive

4.4. Processing standard C

- Only basic checks are made
- Data remains in the format in which it was received
- A basic catalogue record only is created
- Non-digital collections are not anonymised or digitised and transferred to another repository.