

Information on the data repository HeiDATA: Dataverse Network™

The “Dataverse Network™ Project” is a web-based software programme for the citation, analysis and publication of research data. It is open-source and has been in development since 2006 by the Institute for Quantitative Social Science (IQSS) at Harvard University.

With the help of this software, the University of Heidelberg is building a long-term repository for research data: HeiDATA. As part of the collaboration with the University of Heidelberg and the Virtual Special Library of Classical Studies “Propylaeum”, DIGITAL CLASSICS ONLINE authors can use the data repository for accompanying or additional material (raw data).

As part of the journal’s open-access policy and for the sake of the proposed long-term archiving, authors are requested to ensure that their data are in open, non-proprietary formats in accordance with the BSI guidelines for long-term archiving.

According to the IT Baseline Protection Catalog (4.170 M) of the German Federal Office for Information Security (BSI), file formats that are used in electronic long-term archiving must meet the following criteria:

- The data format should have possible long-term relevance.
- The document structure should be unambiguous.
- The document content should be processed electronically.
- Legal requirements must be taken into account.
- Grammar and semantics of the data format should be documented so that subsequent migration to other formats is possible.
- Characteristics of the original document (electronic or paper) should be able to be clearly detected later, even if the original document is no longer available.

These requirements are met by following file types and document markup languages:

A. Structure formats: PDF/A, SGML, XML. When archiving in SGML or XML, however, semantics specification is necessary (DTD - Document Type Definition) and possibly also the layout information to file in XSL or as an additional TIFF document.

B. Image formats: TIFF and JPEG. For archiving it is recommended to choose a lossless compression.

C. Audio and video formats: The digital processing of audio and video data already incurs large amounts of data, so the suitability of the lossless compression method has to be checked with regard to the specific application.

In addition, for two-dimensional data tables we recommend UTF-8 encoded plain text documents with delimiter-separated values, i.e. CSV (comma-separated values) or TSV (tab-separated values).