

# Which kind of data can we preserve?

## Preservation Levels and digital preservation measures in the ETH Data Archive

In order to keep data accessible and usable over the long term, the use of open, documented and widely used standard file formats is an essential prerequisite. Please find information about our assessment of various file formats on the page "[File formats for archiving](#)". However, the suggestions made there are often not applicable in the context of research data. Therefore, the ETH Data Archive will basically accept data in any file format. As a consequence, in many cases only bitstream preservation, i.e. storage of the data in its original form, can be guaranteed. Maintaining the usability of data in future IT environments is an ambitious task which can only be achieved for a limited number of well documented and non-proprietary file formats.

The services we provide for long-term preservation of the data stored in the ETH Data Archive therefore depend on the suitability of the file format for long-term preservation rather than on the retention period chosen.

### 1. Format classifications

#### ***Not suitable for archiving – Bitstream Preservation only***

For data in proprietary or not clearly identifiable formats (for example experimental data in self-defined formats or output of a measuring instrument in a vendor's proprietary format) we guarantee an undamaged storage in their original form (bitstream preservation). We cannot carry out any active preservation measures beyond this. As a rule, such data are suitable for a limited retention period. The same is true for encrypted or otherwise protected data, even if its file format would actually allow active preservation measures.

#### ***Suitable to only a limited extent – Limited Preservation Support***

For some formats, we are monitoring their development and possible risks and may be able to carry out file format migrations in some cases.. Nevertheless, it is difficult to estimate and assess the consequences that may arise from such file format migrations. This is usually the case for proprietary formats that are widely used (e.g. Microsoft Office Open XML). The opportunities for action are therefore limited and extensive preservation measures cannot be guaranteed.

The unchanged storage (Bitstream Preservation) is guaranteed in any case.

#### ***Recommended – Full Preservation Support***

*On this level, reasonable precautions will be taken in order to maintain the data's usability over the long term.*

Such preservation measures can only be taken for formats that are widely used and follow open, documented standards.

Whenever possible, we recommend choosing a standard format for archiving of precious, non-reproducible data such as long-time observational data and series of measurements.

Again, the unchanged storage (Bitstream Preservation), is an essential prerequisite for the maintenance of usability.

For the assignment of particular file formats or groups of formats into the three classes, please refer to our [format recommendations](#). For formats that are not listed, we will not conduct any active preservation measures. The format recommendations will be updated regularly. The current format recommendation also applies to files that are already stored in the ETH Data Archive, regardless of the format recommendation that was active by the time of their upload.

### 2. Preservation measures

Measure	Full Preservation Support	Limited Preservation Support	Bitstream Preservation only
<b>Format identification</b> An unambiguous identification of the file format is an inevitable prerequisite for all active preservation measures. If a format cannot be identified, it can only be preserved as is (Bitstream Preservation).	x	x	x
<b>Extraction and storage of technical metadata</b> Depending on the availability of tools for each format	(x)	(x)	(x)
<b>Validation against established schemes or specifications</b> Depending on the availability of tools for each format If needed, errors can be corrected before the ingest.	(x)	(x)	(x)
<b>Format specific risk analysis</b> Risks can be inherent to the file format or can be gathered from specific properties in the technical metadata.	x	x	

<p><b>Conversion before ingest</b></p> <p>Will be offered in the framework of defined archiving workflows and can (partly) be automated. As a rule, the original format as well as the result of the conversion will be archived.</p>		x	
<p><b>Periodic reanalysis of archived data</b></p> <p>Can new formats be identified?</p> <p>Can additional technical metadata be analyzed with new technical extractors?</p>	(x)	x	x
<p><b>Strategic Monitoring of identified formats</b></p> <p>Has the risk situation changed, is obsolescence to be feared or does any successor format exist that should be evaluated, in order to potentially convert data?</p>	x	x	
<p><b>Migration</b></p> <p>Conversion of the original format into a newer format and archiving as a new version of the object. The former version of the object can be reconstructed anytime.</p>	x	(x)	