Data preservation at RODA

Data preservation and data protection are the most important objectives at RODA. Long term data preservation and reuse of digital data is not possible without a full set of measures, carefully defined and monitored, which refer to both data archival procedures and file formats used for this purpose.

Social research data is unusable without the accompanying metadata. This is why in RODA terminology, the concept of “data” refers to the ensemble of metadata + data.

Data preservation and dissemination

Data preservation refers to all systems and procedures trying to make sure that the data will be readable and usable in the foreseeable future and beyond. That involves choosing certain data storage formats, changing those formats as needed and monitoring the current usage patterns.

Storing data for preservation goes often against storing data for dissemination.

For example, many of the social researchers use SPSS for data processing, statistical analysis and generally to store their studies and variables. SPSS is one of the most enduring software available (first version appeared in 1968) and it can be reasonably expected that it will be available for some foreseeable future. However, SPSS is a binary format and if at some point in the future no software is available to process it, the data will be lost.

There are several types of data, from digital format point of view:

- Alphanumeric data (variable description, study description, people names, etc)
- Numeric data (values, date/time)
- Images
- Video
- Audio

RODA employs two data preservation strategies:

- storing data in a fully open format (e.g. ASCII text or at least UTF8 test). This is not possible with images or videos
- periodic migration to available formats. Currently, images are stored as PNG and JPG, because these are web formats that every computer can process. In the future, if these formats are replaced with new ones, the images from the archive will be migrated to the new formats. In this case, the storage format is the same as the dissemination one.

There are two types of migration: format migration and version migration. Format migration implies that the data format is changed completely (e.g. from PNG to TIFF) whereas version migration means that the format is kept, but it is updated as it evolves (e.g. from TIFF 5.0 to TIFF 6.0).

The original data, with its original file formats will be also preserved.

File migration between formats will be done as needed, according to technology monitoring and file format popularity and accessibility. The decision to migrate one or more files to a different format can be taken with storage (data preservation) or dissemination but is usually determined by one of the following reasons:

- version change (a new version of the current format is available). There are many ways in which a format version can change and sometimes the new version may not be suitable for the data any more. If that is the case, we may decide to maintain the old version (as long as it can still be used) or to migrate to a different format.
- format obsolescence (a certain file format is no longer used or it is used by very few people)
- another format that can hold the same data becomes more attractive for preserving or disseminating the data.

In rare cases, migrating to a new version or format can lead to data alteration or data loss. RODA carefully checks these cases and if a recommended migration path will have such a consequence another option is searched for.

Data protection

Data protection involves all steps taken to assure that the data maintains its original content and is only altered when needed (to be corrected and/or enhanced). All corrections should be monitored and logged and the original state of the data should always be available for inspection (no overwrite).

Alteration of alphanumeric data is easily monitored and tracked. All modifications to data alphanumeric data will be stored in audit tables, together with date and time and person responsible.

Alteration of binary data (images, video) is hard to track but, is easily observed. All binary files will be tracked using hash checksums and thus, easily validated.

Technical details

RODA will maintain the data in two storage systems:

- SQL database, that is needed for fine grained operations through the web interface

- XML data files using DDI format for communication with similar systems

Both SQL database and XML (and other associated files such as images, video, etc.) will be stored on RAID Arrays which allows rapid recovery from disk failure.

All data will be backed up daily to a backup server and further backed up on tape.

No compression will be used. Using compression on a text file means transforming a perfectly readable text file in a closed binary file susceptible to corruption.