Data formats for preservation:

What you need to know when creating a DMP

THE CONTEXT

Different types of data are acquired, processed and stored (preserved and/or archived) in different ways and can be discipline specific. When starting a new project and creating a DMP, one of the first considerations to make should be to decide, in advance, which file formats to use. Many proprietary file formats are "containers" for standard file formats. By packaging them into these containers, a software and/or hardware developer can provide additional functionality, usually by streamlining a process, to analyse data acquired on their platform. However, this has the negative consequence of making these data less interoperable.

Moreover, file formats can be either lossless or lossy: that is, whether data is uncompressed (such as TIFF for images) or compressed (such as JPEG for images) to remove redundant information and thus reduce file size. It is common practice to do analyses on lossy data but this does not necessarily mean that these data should be the ones that should be kept for long-term storage. In this context, it is highly likely that the most important file to consider for long-term storage through its curation lifecycle is either the first file (that which was initially captured from an instrument) or a direct lossless standard file format version from this one (see also guide on raw data).

For H2020 projects, the inclusion of data management plans is default requirement, according to a document issued by the European Commission:

- "H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020" - http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

WHY IS IT NECESSARY?

The key consideration to make is longevity and interoperability and to be as FAIR as much as possible - proprietary versus standard formats: equipment used, brand of software/hardware and research discipline may all be contributing factors. There is no guarantee that existing proprietary file formats will exist in the future. For example, many Microsoft file formats such as Word and Excel may be in common use now, but this does not negate the possibility that they will become obsolete in the future. Software and file format obsolescence will be even more pronounced for bespoke file formats that were created as part of an individual project.

HOW TO DEAL WITH THIS?

As an example, in the biomedical imaging field, a realisation of the huge variety of file formats that exist led to an initiative to make these interoperable. As part of the OMERO project, Bioformats is a software plugin, which allows the conversion of multiple established proprietary and standard file formats. Image analysis software such as ImageJ (free and open source) has adopted Bioformats as a plugin to allow users to read and write their image data without having to consider their origin. However, such tools may not always exist for different disciplines, and a researcher should consider storing their acquired data in a standard format at the earliest available opportunity. Many (most?) commercial and open source software packages allow conversion of data into standard formats and this should be exploited.
During the course of the digital revolution, a number of file formats have been recognised to be the file formats of choice for longevity and interoperability.

Please find below some useful links to resources about data formats for long-term storage:

- 4TU Center for Research Data - “data description and formats” - https://researchdata.4tu.nl/en/publishing-research/data-description-and-formats/
- DANS | Data Arching and Networked Services - “file formats” - https://dans.knaw.nl/en/deposit/information-about-depositing-data/before-depositing/file-formats

As an example, the following table describes a variety of file formats for different disciplines that are either recommended or acceptable (from the UK Data Service):

Type of data	Recommended formats	Acceptable formats
Tabular data with extensive metadata **variable labels, code labels, and defined missing values**	SPSS portable format (.por) delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) structured text or mark-up file of metadata information, e.g. DDI XML file	proprietary formats of statistical packages: SPSS (.sav), Stata (.dta), MS Access (.mdb/.accdb)
Tabular data with minimal metadata column headings, variable names	comma-separated values (.csv) tab-delimited file (.tab) delimited text with SQL data definition statements	delimited text (.txt) with characters not present in data used as delimiters widely-used formats: MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf), OpenDocument Spreadsheet (.ods)

Type of data	Recommended formats	Acceptable formats
Geospatial data vector and raster data	ESRI Shapefile (.shp, .shx, .dbf, .prj, .sbx, .sbn optional) geo-referenced TIFF (.tif, .tfw) CAD data (.dwg) tabular GIS attribute data Geography Markup Language (.gml)	ESRI Geodatabase format (.mdb) MapInfo Interchange Format (.mif) for vector data Keyhole Mark-up Language (.kml) Adobe Illustrator (.ai), CAD data (.dxf or .svg) binary formats of GIS and CAD packages
Textual data	Rich Text Format (.rtf) plain text, ASCII (.txt) eXtensible Mark-up Language (.xml) text according to an appropriate Document Type Definition (DTD) or schema	Hypertext Mark-up Language (.html) widely-used formats: MS Word (.doc/.docx) some software-specific formats: NUD*IST, NVivo and ATLAS.ti
Image data	TIFF 6.0 uncompressed (.tif)	JPEG (.jpeg, .jpg, .jp2) if original created in this format GIF (.gif) TIFF other versions (.tif, .tiff) RAW image format (.raw) Photoshop files (.psd) BMP (.bmp) PNG (.png) Adobe Portable Document Format (PDF/A, PDF) (.pdf)
Audio data	Free Lossless Audio Codec (FLAC) (.flac)	MPEG-1 Audio Layer 3 (.mp3) if original created in this format Audio Interchange File Format (.aif) Waveform Audio Format (.wav)

Type of data	Recommended formats	Acceptable formats
Video data	MPEG-4 (.mp4) OGG video (.ogv, .ogg) motion JPEG 2000 (.mj2)	AVCHD video (.avchd)
Documentation and scripts	Rich Text Format (.rtf) PDF/UA, PDF/A or PDF (.pdf) XHTML or HTML (.xhtml, .htm) OpenDocument Text (.odt)	plain text (.txt) widely-used formats: MS Word (.doc/.docx), MS Excel (.xls/.xlsx) XML marked-up text (.xml) according to an appropriate DTD or schema, e.g. XHMTL 1.0

When writing a DMP, researchers are advised to refer to tables such as this to help decide the best file formats to use for their project and to state this clearly.

Video

Utrecht University. "*Preserving research data in the optimal, technically correct way*" (How to minimize the risk of losing data. Here you'll learn which methods there are to preserve your research data in an optimal way) - https://www.youtube.com/watch?v=qENaO0Lk6eo#action=share

RESOURCES USED IN THIS GUIDE

Publications

Scaled and automated preservation planning for highly diverse digital collections: the Integrated Preservation Suite.
DOI=https://dx.doi.org/10.7207/twr14-02
https://zenodo.org/record/1299966/preview/ipres_ips_revised.pdf

Science Europe Guidance Document: Presenting a Framework for Discipline-specific Research Data Management (January 2018) https://www.scienceeurope.org/wp-content/uploads/2018/01/SE_Guidance_Document_RDMPs.pdf

The FAIR Guiding Principles for scientific data management and stewardship | US National Library of Medicine National Institutes of Health
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/

Other resources

Best practice for file formats (Stanford Libraries)
https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats

Consortium of European Social Sciences Data Archives (CESSDA)
Adapt your Data Management Plan: a list of Data management questions based on the Expert Tour Guide on Data management
https://www.cessda.eu/content/download/3844/35033/file/20171117DMPQuestionsCESSDAExpertTourGuide.pdf

Consortium of European Social Sciences Data Archives (CESSDA)
Expert Tour Guide on Data Management (on section 3 - Process | File formats and data conversion
https://www.cessda.eu/Research-Infrastructure/Training/Expert-Tour-Guide-on-Data-Management

Data management knowledge, tools, and training | DTL - Dutch Techcenter for Life Sciences
https://www.dtls.nl/fair-data/research-data-management/data-management-knowledge-tools/

Digital Preservation Coalition - Digital Preservation Handbook
https://www.dpconline.org/handbook

FAIR data: what it means, how we achieve it, and the role of RDA (presentation from Sarah Jones)
https://pt.slideshare.net/sjDCC/fair-data-what-it-means-how-we-achieve-it-and-the-role-of-rda

Storing and Preserving Data (Utrecht University) https://www.uu.nl/en/research/research-data-management/guides/storing-and-preserving-data

The National Archives:
http://www.nationalarchives.gov.uk/information-management/manage-information/digital-records-transfer/file-formats-transfer/

UK Data Service guidelines:
https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats

University of Vienna (format informations and supported formats)
https://datamanagement.univie.ac.at/en/about-phaidra/formats/

University of Virginia Library
https://data.library.virginia.edu/data-management/plan/format-types/