**What types of data are we talking about?**

**Data** can mean many different things, and there are many ways to classify it.  Two of the more common are:

* Primary and Secondary: Primary data is data that you collect or generate.  Secondary data is created by other researchers, and could be their primary data, or the data resulting from their research.
* Qualitative and Quantitative: Qualitative refers to text, images, video, sound recordings, observations, etc.  Quantitative refers to numerical data.

There are typically five main categories that it can be sorted into for data management purposes. The category that you choose will then have an effect upon the choices that you make throughout the rest of your data management plan.

**Observational**

* Captured *in situ* (real-time)
* Cannot be recaptured, recreated, or replaced (unique data).
* Examples include sensor readings, telemetry, survey results, images, and human observation.

**Experimental**

* Data from lab equipment and under controlled conditions
* Often reproducible, but can be expensive to do so
* Examples include gene sequences, chromatograms, magnetic field readings, and spectroscopy.

**Simulation**

* Data generated from test models studying actual or theoretical systems
* Models and metadata where the input more important than the output data
* Examples include climate models, economic models, and systems engineering.

**Derived or compiled**

* The results of data analysis, or aggregated from multiple sources
* Reproducible (but very expensive)
* Examples include text and data mining, compiled database, systems engineering, and 3D models

**Reference or canonical**

* Fixed or organic collection datasets, usually peer-reviewed, and often published and curated
* Examples include gene sequence databanks, census data, chemical structures.

Data can come in many forms. Some common ones are text, numeric, multimedia, models, audio, code, software, discipline specific (i.e., FITS in astronomy, CIF in chemistry), video, and instrument.

**What are the issues around file formats?**

**File Formats** should be chosen to ensure sharing, long-term access and preservation of your data.  Choose open standards and formats that are easy to reuse.  If you are using a different format during the collection and analysis phases of your research, be sure to include information in your documentation about features that may be lost when the files are migrated to their sharing and preservation format, as well as any specific software that will be necessary to view or work with the data.

**Best practice for file format selection include:**

* non-proprietary – open, documented standard
* unencrypted – neither full or partial
* uncompressed – no “lossy” or proprietary compression
* no embedded files, programs, or scripts
* no password protection
* use common character encodings – ASCII, Unicode, UTF-8

Remember to retain your original unedited raw data in its native formats as your source data.  Do not alter or edit it.  Document the tools, instruments, or software used in its creation.  Make a copy of it prior to any analysis or data manipulations.

**Recommended Digital Data Formats:**

Text, Documentation, Scripts: XML, PDF/A, HTML, Plain Text.

Still Image: TIFF, JPEG 2000, PNG, JPEG/JFIF, DNG (digital negative), BMP, GIF.

Geospatial: Shapefile (SHP, DBF, SHX), GeoTIFF, NetCDF.

Graphic Image:

* raster formats: TIFF, JPEG2000, PNG, JPEG/JFIF, DNG, BMP, GIF.
* vector formats: Scalable vector graphics, AutoCAD Drawing Interchange Format, Encapsulated Postscripts, Shape files.
* cartographic: Most complete data, GeoTIFF, GeoPDF, GeoJPEG2000, Shapefile.

Audio: WAVE, AIFF, MP3, MXF, FLAC.

Video: MOV, MPEG-4, AVI, MXF.

Database: XML, CSV, TAB.

Adapted: Library of Congress [Recommended Formats Statement](http://www.loc.gov/preservation/resources/rfs/stillimg.html#photodigital), [UK Data Archive](http://data-archive.ac.uk/create-manage/format/formats-table), [Univ. Virginia Library](https://data.library.virginia.edu/data-management/plan/format-types/)