12 Principles of Data Management

Plato Smith, UF Libraries

1. Define contents of your data files
   - Add comments in code (i.e. this block of code does this, that, or method exception)
2. Define the variables
   - Develop codebook (i.e. understood by other programmers/researchers)
3. Use consistent data organization
   - Define folder and file names and structure – and use them consistently
   - Limit the depth of subfolders to no more than 2
4. Use stable file formats
   - Include other file formats, if possible, than original (i.e. VCF (Variant Call Format) saved as CSV (Common Separated Value) for increased sharing and use)
5. Assign descriptive file names
   - Use unique names
   - Avoid spaces (use underscore instead of spaces)
   - Use ASCII Characters only
   - Document, share, and evaluate
   - Separate classes of products: raw data, derived data, graphics, code, documents, etc.
6. Preserve processing information
   - Automate whenever possible (See: attached revise-SIGMOD_Reproduciblity_Template)
     - Source for template: http://db-reproducibility.seas.harvard.edu/ and http://daslab.seas.harvard.edu/
7. Perform basic quality assurance
   - Clean and Create (scripts)
     - Data Cleaning Process
     - Data cleansing for improvement
     - Data change management
8. Provide documentation
   - Provide metadata for code (See: Reproducibility template)
   - Use comments within code to describe code functions
9. De-identify your data
   - Strip personal identifying information (PID)
     - See 18 HIPAA Identifiers on attached Zenodo Brief Intro and Demo presentation)
     - Ensure compliance (i.e. See HIPAA De-identification methods - https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard)
10. Protect your data
    - Initiate access restrictions, if necessary (e.g. open, closed, restricted, no access)
    - Utilize secure computing environment, if necessary (i.e. ResearchVault - https://www.rc.ufl.edu/services/restricted-data/researchvault/)
11. Preserve your data
    - Archive GitHub, code in Zenodo
    - Practice with Zenodo sandbox first before using Zenodo production (recommended)
      - Zenodo sandbox - https://sandbox.zenodo.org/
12. Cite your data
    - Make Your Code Citable - https://guides.github.com/activities/citable-code/
    - Make your data Findable, Accessible, Interoperable, and Reusable (FAIR) – a resource: https://www.ands.org.au/working-with-data/fairdata
    - Citing Data (an example - NCSU Libraries, https://www.ands.org.au/working-with-data/fairdata)