2016

Preparing Data for Sharing and Archiving

Data Management Services, Zach S. Henderson Library

Preparing Data for Sharing and Archiving

Significant portions of this presentation are adapted from the Cornell University Research Data Management Services Group website under a Creative Commons Attribution 4.0 International License, and from ICPSR's Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle (2012, 5th ed.). Ann Arbor, MI.

Agenda

- Why share and archive data?
- What should I share and archive?
- Data collection, file creation, and management
- Metadata creation
- Protecting subjects
- Copyright and re-use licensing

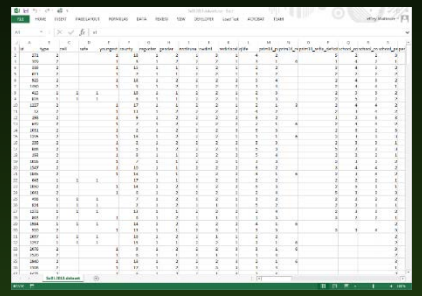
Why share and archive data?

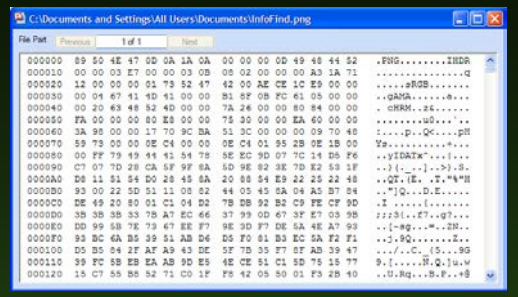
- Many *research funders* now require PIs to maximize open public access to data products.
- Many *publishers* now require open access to replication data as a condition of publication.
- It benefits *you*, your *collaborators*, and your *research community*.

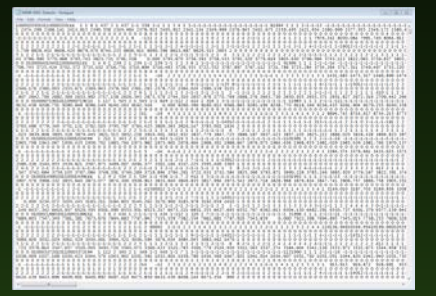
What should I share and archive?

Content + Metadata

- Content = The *bit-level data*, the "target" of preservation.



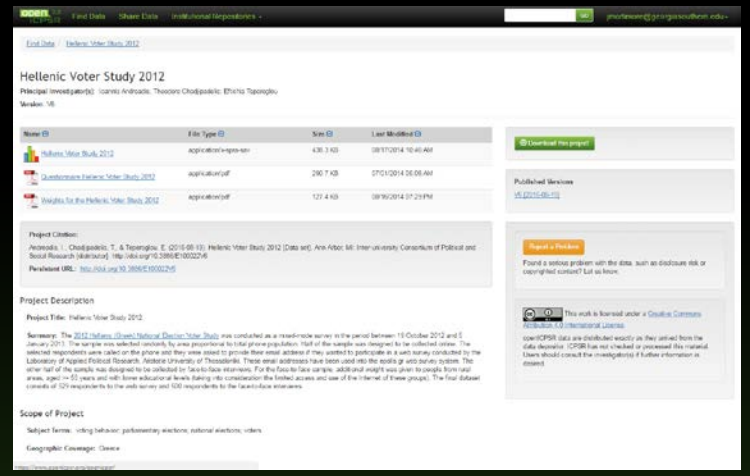




- Metadata = The *information needed* to find, access, understand, use, and preserve the content.

Metadata

- Descriptive Information
  Makes the data discoverable
- Packaging Information
  Explains how the data is organized
- Representation Information
  Makes the data understandable; renders the bit-level content into something meaningful.
- Preservation Information
  Provides information to support long-term preservation and use



Data collection, file creation, and management

- Data and file structure
  What is the data file going to look like and how will it be organized? What file types will be used?
- Naming conventions
  How will files and variables be named? What conventions will be used?
- Data integrity
  How will the data be captured and checked for accuracy and integrity? How will versions be checked?
- Dataset documentation
  When and how will documentation be produced? What standards should be used/applied to make it usable by others?
- Variable construction
  What variables will be constructed following data collection? According to what standards and how will they be documented?
- Project documentation
  How will decisions be documented over the course of the research (e.g., field procedures, coding decisions, variable construction)?

Data collection, file creation, and management

The likelihood of long-term preservation of content and functionality is higher when file formats possess the following characteristics:

- complete and open documentation
- platform-independence
- non-proprietary (vendor-independent)
- no “lossy” or proprietary compression
- no embedded files, programs or scripts
- no full or partial encryption
- no password protection
- uncompiled

Table of Recommended File Formats for Long-Term Data Curation

Content Type	High probability for long-term preservation	Medium probability for long-term preservation	Low probability for long-term preservation
Text	• Plain text (encoding: USASCII, UTF-8, UTF-16 with BOM) • XML (includes XSD/XSL/XHTML, etc.; with included or accessible schema) • PDF/A-1 (ISO 19005-1) (*.pdf)	• Cascading Style Sheets (*.css) • DTD (*.dtd) • Plain text (ISO 8859-1 encoding) • PDF (*.pdf) (embedded fonts) • Rich Text Format 1.x (*.rtf) • HTML (include a DOCTYPE declaration) • SGML (*.sgml) • Open Office (*.sxw/*.odt) • OOXML (ISO/IEC DIS 29500) (*.docx)	• PDF (*.pdf) (encrypted) • Microsoft Word (*.doc) • WordPerfect (*.wpd) • DVI (*.dvi) • All other text formats not listed here
Raster Image	• TIFF (uncompressed) • JPEG2000 (lossless) (*.jp2) • PNG (*.png)	• BMP (*.bmp) • JPEG/JFIF (*.jpg) • JPEG2000 (lossy) (*.jp2) • TIFF (compressed) • GIF (*.gif) • Digital Negative DNG (*.dng)	• MrSID (*.sid) • TIFF (in Planar format) • FlashPix (*.fpx) • PhotoShop (*.psd) • RAW • JPEG 2000 Part 2 (*.jpf, *.jpx) • All other raster image formats not listed here
Vector Graphics	• SVG (no Java script binding) (*.svg)	• Computer Graphic Metafile (CGM, WebCGM) (*.cgm)	• Encapsulated Postscript (EPS) • Macromedia Flash (*.swf) • All other vector image formats not listed here
Audio	• AIFF (PCM) (*.aif, *.aiff) • WAV (PCM) (*.wav)	• SUN Audio (uncompressed) (*.au) • Standard MIDI (*.mid, *.midi) • Ogg Vorbis (*.ogg) • Free Lossless Audio Codec (*.flac) • Advance Audio Coding (*.mp4, *.m4a, *.aac) • MP3 (MPEG-1/2, Layer 3) (*.mp3)	• AIFC (compressed) (*.aifc) • NeXT SND (*.snd) • RealNetworks 'Real Audio' (*.ra, *.rm, *.ram) • Windows Media Audio (*.wma) • Protected AAC (*.m4p) • WAV (compressed) (*.wav) • All other audio formats not listed here

Data collection, file creation, and management

- Make file names unique, including the most important identifying information. Elements of a good file name may include:
  - project name, acronym, or research data name
  - study title
  - location information
  - researcher initials
  - date (consistently formatted, i.e. YYYYMMDD)
  - version number
- Use leading zeros to enable sorting ("1-100" should be numbered 001-100).
- Use underscores to separate elements; avoid special characters, spaces, and periods.
- File names should be short enough to be readable, while still being meaningful. For example:

DryValleySoil_ICPOES_20101115_JDS.dat

DryValleySoil is the project name, ICPOES is the instrument from which the data originated, 20101115 is the date of the sample run on the instrument, and JDS are the initials of Jane Doe Scientist.

Data collection, file creation, and management

- Keep track of versions when working with data.
- Some research management tools, like Electronic Lab Notebooks (ELN), Open Science Framework, and Box support version control. Other options include using a naming scheme or version control software.
- Best practices include:
  - Save an untouched copy of the raw data and leave it that way. Always work on copies of the "safe" untouched copy.
  - Avoid ambiguous labels, such as 'revision', 'final', 'final2', etc. Instead, use a file naming convention (like v001, v002 or v1_0, v1_2, v2_0).
  - Use a directory structure naming convention that includes version information.
  - Consider using version control software (e.g., GitHub). However, test any versioning tool to make sure that it supports retention of and/or reversion to previous versions.

Data collection, file creation, and management

- Document and Use Directory Structure Naming Conventions
- Directory top-level folders should include the project title, unique identifier, and date (year), but the files themselves should be well-described independent of the directory structure.
- Consider creating a brief description of the contents of major folders and providing an overview of the directory structure. This can be a text document or readme file that is stored in a top-level folder or shared space.
- Provide enough information to help someone else understand the contents and organization of your files in your absence.

Data collection, file creation, and management

- When preparing tabular data for description and archiving:
  - Only include the data; do not include figures or analyses.
  - Consider aggregating data into fewer, larger files, rather than many small ones. Compress cautiously.
  - If a repository has no file format requirements, consider tab- or comma-delimited text (*.txt or *.csv).
  - Column headings should be meaningful, but not overly long.
  - Use only alphanumeric characters, underscores, or hyphens in column headings.
  - Some programs expect the first character to be a letter; start column headings with a letter.
  - Indicate units of measurement in column headings and specify units in the metadata.
  - Use only the first row to identify a column heading.
  - Examples of good column headings:

max_temp_celsius - not max temp celsius (includes spaces)
airport_faa_code - not airport/faa code (includes special characters)

Data collection, file creation, and management

- Use standard codes or names when possible. Examples include using Federal Information Processing (FIPS) codes for geographic entities and the Integrated Taxonomic Information System (ITIS) for authoritative species names.
- When using non-standard codes, an alternative to defining the codes in the metadata is to create a supplemental table with code definitions (e.g., in a codebook, the instrument, or a "readme" file).
- Avoid using special characters such as commas, semicolons, or tabs, in the data itself if the data file is in (or will be exported to) a delimited format.
- Do not rely on special formatting that is available in spreadsheet programs, such as Excel.
- Indicate date information in an appropriate machine-readable format, such as yyyymmdd or yyyy-mm-dd. Indicate time zone (including daylight savings, if relevant) and use of 12-hour or 24-hour notation in the metadata.
- Alternately, use the ISO standard for formatting date and time strings. The standard accommodates time zone information and uses 24-hour notation.
- Use a standard method to identify missing data; do not use zeroes or leave a cell blank. -999 or -9999 is a common convention. Indicate the code for missing data in the metadata.

Data collection, file creation, and management

- Consider performing basic data quality assurance to detect errors or inconsistencies in data. Some common techniques are:
  - Spot check some values in the data to ensure accuracy.
  - Enter data twice and compare both versions to catch errors.
  - Sort data by different fields to spot outliers and empty cells.
  - Calculate summary statistics, or plot data to catch erroneous or extreme values.

- Provide summary information about the data and include it in the metadata. For example:
  - number of columns
  - max, min, or mean of parameters
  - number of missing values
  - total file size

Metadata creation

- Metadata is documentation that describes data.
- In a lab setting, much of the content used to describe data is initially collected in a notebook; metadata is a more formal, sharable expression of this information.
- Metadata can include content such as contact information, geographic locations, details about units of measure, abbreviations or codes used in the dataset, instrument and protocol information, survey tool details, provenance and version information and much more.
- Metadata can take many different forms, from free text to standardized, structured, machine-readable, extensible content.
- Specific disciplines, repositories or data centers may guide or even dictate the content and format of metadata.

Metadata creation

Examples metadata standards:

- Dublin Core - domain agnostic, basic and widely used metadata standard
- DDI (Data Documentation Initiative) - common standard for social, behavioral and economic sciences, including survey data
- EML (Ecological Metadata Language) - specific for ecology disciplines
- ISO 19115 and FGDC-CSDGM (Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata) - for describing geospatial information
- MINSEQE (MINimal information about high throughput SEQeuencing Experiments) - Genomics standard
- FITS (Flexible Image Transport System) - Astronomy digital file standard that includes structured, embedded metadata

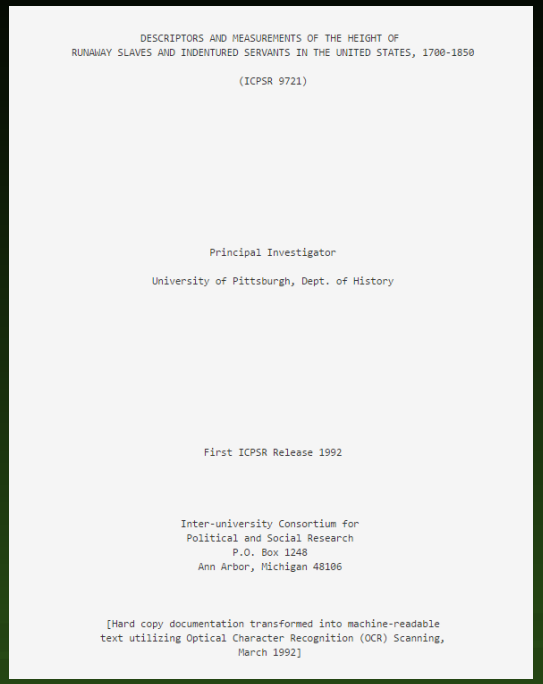
Metadata creation

Important metadata elements may include:

- Principal investigator(s) [Dublin Core -- Creator]
- Title [Dublin Core -- Title]
- Funding sources
- Data collector/producer
- Project description [Dublin Core -- Description]
- Sample and sampling procedures
- Weighting
- Substantive, temporal, and geographic coverage [Dublin Core -- Coverage]
- Data source(s) [Dublin Core -- Source]
- Unit(s) of analysis/observation
- Variables
- Related publications
- Technical information on files
- Data collection instruments
- Flowchart of the data collection instrument
- Index or table of contents
- List of abbreviations and other conventions
- Interviewer guide
- Coding instrument

Metadata creation

Metadata typically is provided via codebooks, XML files (e.g., DDI), manifests, etc.



DESCRIPTORS AND MEASUREMENTS OF THE HEIGHT OF
RUNAWAY SLAVES AND INDENTURED SERVANTS IN THE UNITED STATES, 1700-1850

(ICPSR 9721)

Principal Investigator

University of Pittsburgh, Dept. of History

First ICPSR Release 1992

Inter-university Consortium for
Political and Social Research
P.O. Box 1248
Ann Arbor, Michigan 48106

[Hard copy documentation transformed into machine-readable
text utilizing Optical Character Recognition (OCR) Scanning,
March 1992]

example DDI encoded metadata

Protecting subjects

Subjects' privacy and confidentiality must be protected consistent with IRB requirements and disciplinary ethics. Variables that endanger privacy and confidentiality include:

- Direct identifiers
  Variables that point explicitly to particular individuals or groups.

- Indirect identifiers
  Variables make unique cases visible when combined with other identifiers.

- Geographic identifiers
  Direct geographic identifiers may include specific addresses. Indirect geographic identifiers may include census tracts, area codes, place of birth or education, etc.

Protecting subjects

Common techniques for treating identifiers to protect subjetcs:

- Removal: Eliminate the variable.
- Top-Coding: Restrict the upper range of a variable
- Collapsing and/or Combining: Combining values into a summary variable
- Sampling: Release a random sample of sufficient size to yield reasonable inferences.
- Swapping: Match unique cases on the indirect identifier, then exchange the values of key variables between the cases. This retains the analytic utility and covariate structure of the dataset while protecting subject confidentiality.
- Disturbing: Add random variation or stochastic error to the variable. This retains the statistical properties between the variable and its covariates, while preventing using the variable as a means for linking records.

Protecting subjects

Alternatives to altering the data:

- Restricted-use datasets
  A dataset released only to approved researchers who agree to abide by rules assuring subjects' privacy and confidentiality is maintained. Researchers are usually given access to the data for a limited time, at the end of which they must return or destroy the data.
- Data Enclaves
  A physical or virtual environment that allows access, but prevents researchers from retaining any of the data.

Copyright and re-use licensing

- Unmediated factual data cannot be copyrighted because it is not possible to copyright facts (e.g., a temperature reading). However:
  - Some data may be protected, such as photographs.
  - Organized data (e.g., a database) has a *thin layer* of copyright protection because of the researcher's creative input into creating it.
- Copyright may govern the use of databases and some kinds original data content, but contract law, trademarks, and other mechanisms are required to regulate factual data.

Copyright and re-use licensing

Creative Commons (http://www.creativecommons.org/) offers a library of standardized licenses, some of which may be used with data. Creative Commons recommends the following three licenses only for data sharing. :

- CC Zero ("CC0")
  Waive all copyright and database rights, including your right to attribution. This license effectively places the database and data into the public domain, and maximizes the likelihood of reuse.
- CC Attribution 4.0 International ("CC BY 4.0")
  Waive all copyright and database rights except the right to attribution. This license balances your right to be acknowledged with encouraging reuse.
- CC Attribution-ShareAlike 4.0 International ("CC BY-SA 4.0")
  Protect your right to attribution, as well as require that any derivative work be shared under the same licensing conditions. This license may result in confusing "license chaining," and discourage some reuse and citation.

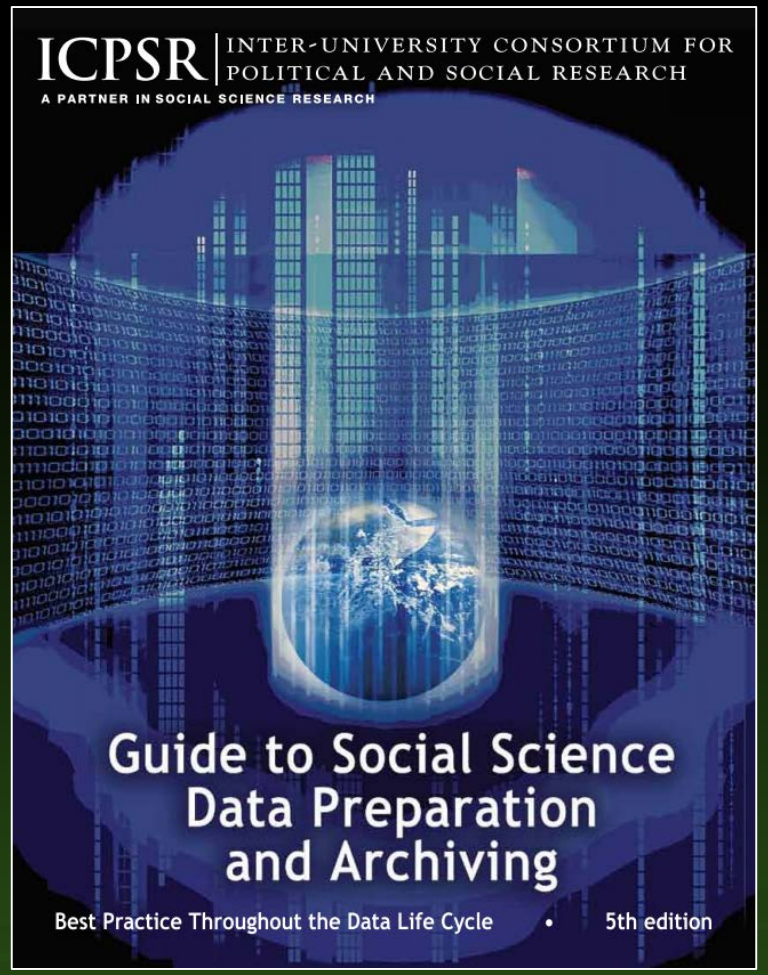
The library recommends using the CC BY 4.0 license in most cases.

Copyright and re-use licensing

Data Ownership at Georgia Southern University:

- Ownership of works produced by Georgia Southern faculty, students, and non-academic staff is governed by the University System of Georgia's Policy on the Use of Copyrighted Works in Education and Research and Georgia Southern University's Intellectual Property and Technology Transfer Policy.
- The precise answer to who owns your data depends on whether the project was created as part of sponsored research; the employment status of the creator; whether the work was conducted pursuant to a specific direction or assigned duty; and, whether substantial university resources were used in the creation of the work.
- Consult with the Office of Research and Economic Development.

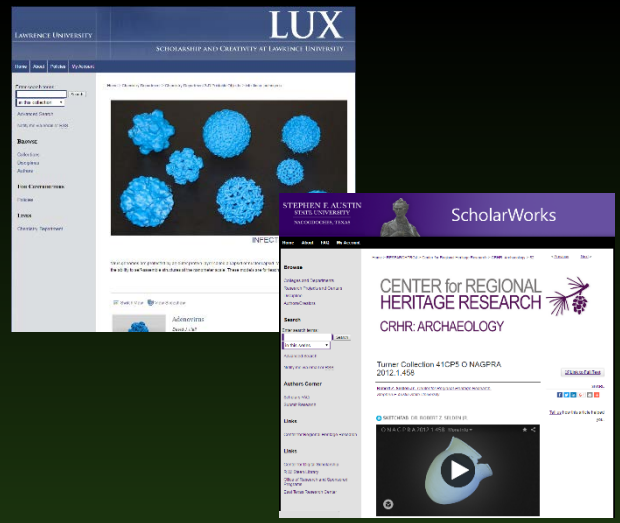
ICPSR Guide to Social Science Data Preparation and Archiving



http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf

Digital Commons for Data

Comprehensive hosted solution for storing, managing, securing, and sharing data:



- Unlimited storage (no additional cost).
- Support for all file types and formats.
- Authorization and access-control tools.
- Fully hosted: no IT resources required.
- Multiple back-ups, cloud storage, and quarterly archives.

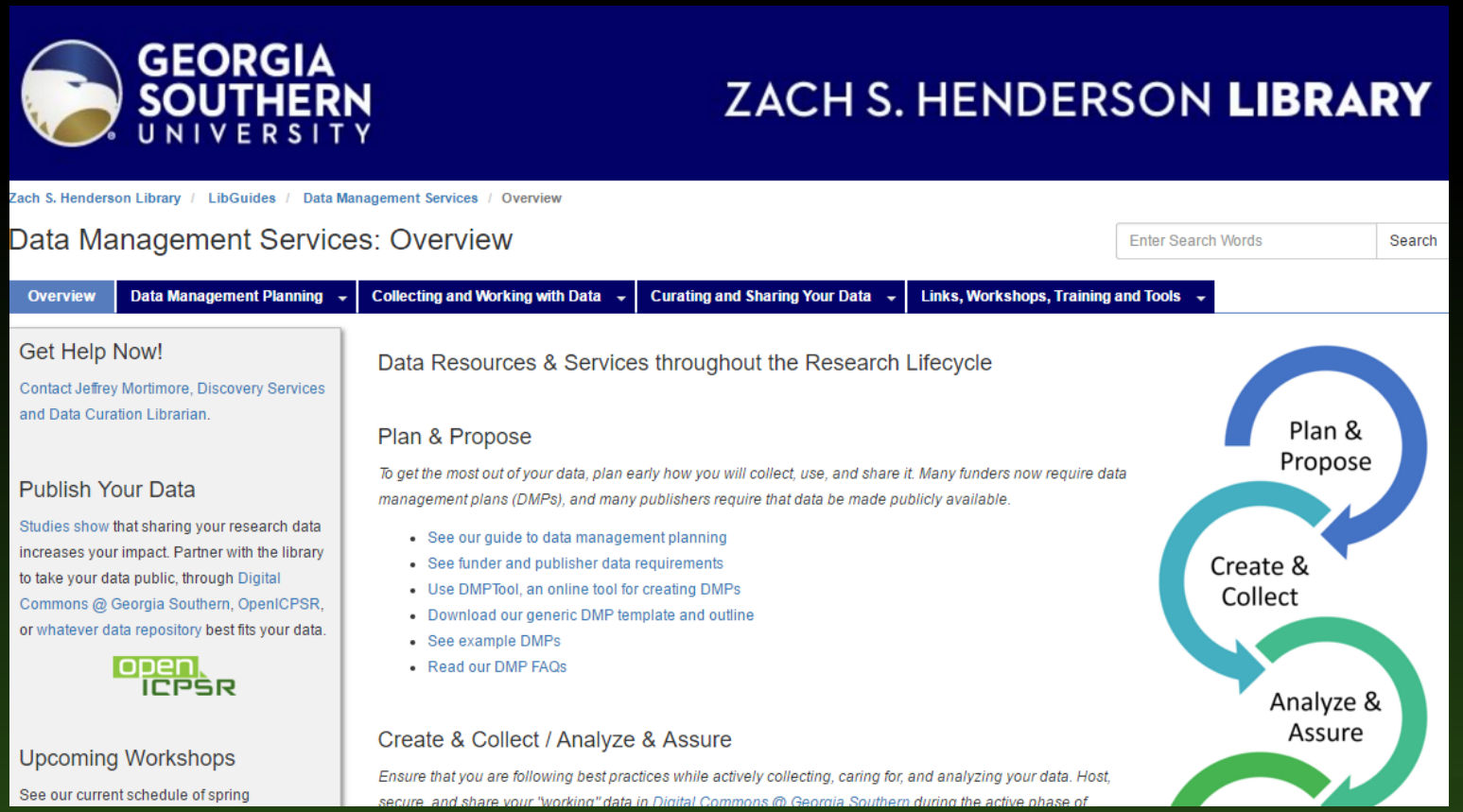
http://digitalcommons.bepress.com/cgi/viewcontent.cgi?article=1000&context=promotional

Archiving and Publication

- Once ready for publication, data objects are re-described, assigned DOIs if needed, and released to the public website.
- Update versions and file types as needed.
- Host on university, departmental, or project-related data structure.
- Backup and Archiving is automatic.
- Links are permanent.
- Link to SelectedWorks profile.

http://georgiasouthern.libguides.com/data

Data Management Services @ Henderson Library



http://georgiasouthern.libguides.com/data