**Group Member: Doreen, Luming**

**Collection Policy for Best Practices of DH in Chinese and Japanese studies Project**

**1. Relevance Criteria for Data Collection**

**Inclusion Criteria:**

- Thematic and Temporal Scope: Data must be related to scholarly contributions on modern and contemporary (post-1900) history, literature, art history, religion, society and thought of China and Japan.

- Academic Value: Data should offer potential for scholarly research, teaching, or broader dissemination within the field of digital humanities.

**Exclusion Criteria:**

- Out-of-Scope Themes: Data and its associated research focusing on periods before 1900 or unrelated to thematic and temporal scope.

- Redundant or Overrepresented Data: Data that does not add unique value or is redundant with existing collections.

- Legal and Ethical Concerns: Data containing sensitive information without proper anonymization, or lacking necessary permissions or rights for sharing.

**2. Minimum Viable Package Requirements (adapted from** [**DRUM policy**](https://conservancy.umn.edu/pages/drum/policies/#data-collection-policy)**)**

Every dataset submitted must include:

- Clear Licensing: Must be explicitly stated, favoring open licenses like CC-BY to enable sharing and reuse.

- Standard or Convertible Formats: Ensuring data is either in a standard format or can be converted to one, to facilitate access and use.

- Comprehensive Documentation: Must include detailed documentation that enables other researchers to understand the context, methodology, and potential applications of the data, facilitating the creation of structured metadata.

**3. Accepted and Non-Accepted Formats (Adapted from** [**Cornell Digital Repository’s file format**](https://guides.library.cornell.edu/ecommons/formats)**)**

**Characteristics of Recommended Formats:**

* Complete and open documentation
* Platform-independence
* Non-proprietary
* No embedded files, programs or scripts
* No full or partial encryption
* No password protection

**Recommended Formats File Formats for Digital Content:**

**Probability for Full Long-Term Preservation**

| Content Type | High Probability | Medium Probability | Low Probability |
| --- | --- | --- | --- |
| 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)  Microsoft Word 2007 or newer (\*.docx) | PDF (\*.pdf) (encrypted)  Microsoft Word 2003 or older (\*.doc)  WordPerfect (\*.wpd)    DVI (\*.dvi)    All other text formats not listed |
| 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 |
| 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 |
| Audio | AIFF (96kHz 16bit PCM) (\*.aif, \*.aiff)  WAV (96kHz 24bit 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 |
| Video | Motion JPEG 2000 (ISO/IEC 15444-4)??\*.mj2)  AVI (uncompressed/native, motion JPEG) (\*.avi)  QuickTime Movie (uncompressed/native, motion JPEG) (\*.mov) | Ogg Theora (\*.ogg)  MPEG-1, MPEG-2 (\*.mpg, \*.mpeg, wrapped in AVI, MOV)  MPEG-4 (H.263, H.264) (\*.mp4, wrapped in AVI, MOV) | AVI (others) (\*.avi)  QuickTime Movie (others) (\*.mov)  RealNetworks 'Real Video' (\*.rv)  Windows Media Video (\*.wmv)  All other video formats not listed |
| Spreadsheet / database | Character delimited text (ASCII or Unicode preferred):  Comma Separated Values (\*.csv)  Delimited Text (\*.txt)  SQL Data Definition Language | OOXML (ISO/IEC DIS 29500) (\*.xlsx)  Excel 2007 or newer (\*.xlsx)  OpenOffice (\*.sxc/\*.ods)  DBF (\*.dbf) | Excel 2003 or older (\*.xls)  All other spreadsheet/ database formats not listed |
| Virtual reality | X3D (\*.x3d) | VRML (\*.wrl, \*.vrml)  U3D (Universal 3D file format) | All other virtual reality formats not listed |
| Computer programs | Uncompiled computer program source code (\*.c, \*.cpp, \*.java, \*.js, \*.jsp, \*.php, \*.pl, \*.py, etc.) |  | Compiled / Executable files (EXE, \*.class, COM, DLL, BIN, DRV, OVL, SYS, PIF) |
| Presentation |  | OpenOffice (\*.sxi/\*.odp)  OOXML (ISO/IEC DIS 29500) (\*.pptx)  PowerPoint 2007 or newer (\*.pptx) | PowerPoint 2003 or older (\*.ppt)  All other presentation formats not listed |

**Non-Accepted Formats:**

- Proprietary or Closed Formats: proprietary database files, or software-specific formats that restrict access and use without specialized tools or are password-protected.

- Legacy Formats: Obsolete digital formats that are no longer supported or commonly used, which pose risks for long-term accessibility.

**4. Size Limitations (Adapted from** [**Cornell Content Policy**](https://guides.library.cornell.edu/ecommons/contentpolicy)**)**

- Maximum Size: Each individual file can be up to 5 GB. For files larger than this limit, submitters must consult with repository administrators to discuss feasibility, potential segmentation, or special accommodations.

**5. Preservation Guarantees**

- Digital Preservation Practices: The repository will ensure that all accepted data formats are preserved using the best practices in digital preservation, which includes:

- Routine Backups: Regularly scheduled backups to prevent data loss.

- Format Migration: Periodic evaluation and migration of data to current formats to ensure ongoing accessibility.

- Longevity Commitment: A minimum of 10 years of data preservation post-submission, with reviews to extend this period as necessary.

- Preservation by Format: Specific strategies will be applied based on the type of data:

- Textual and Tabular Data: Preserved in their native or converted open formats with checksum verification and redundancy.

- Images: Stored in high-quality formats with metadata detailing preservation status and history.

**DH Categories and Their Importance** (adapted from [a submission guideline in East Asian studies](https://airtable.com/app3bRhdc6N5x4Jrm/shrKMyt963WcaXTGz))

To enhance the repository's usability, submissions can be categorized by the digital methods used and the project’s content. These broad categories simplify data organization and retrieval, helping users find relevant resources easily. Contributors are encouraged to select multiple categories to accurately reflect their project's scope and methodology. We invite proposals for new categories to ensure our classifications encourage innovation and inclusivity.

Explanation of Categories

DH Methods and Scope

1. Text Mining:

- Projects that involve deriving or analyzing unstructured data from texts, using methods such as statistical pattern learning or natural language processing.

2. Network Analysis:

- Includes projects focused on creating or analyzing networks, including network visualizations like graphs or node-link diagrams that express relationships.

3. OCR (Optical Character Recognition):

- Projects that convert images of typed, handwritten, or printed text into machine-encoded text through mechanical or electronic means.

4. Visual Tools:

- Projects that involve the exploration and analysis of visual sources, such as art or digitized primary texts, featuring overlays, translations, or annotations.

5. Maps/Mapping and GIS:

- Projects that produce and analyze maps or visualize data on maps.

6. Data Set/Statistics:

- Projects that provide data sets or statistics for public use, aimed at supporting empirical research or public inquiry.

7. Topic Modeling:

- Projects using statistical models to discover the abstract "topics" that occur within a collection of documents.

8. Archive:

- Collections of documents or materials that serve as historical records.

9. Database:

- Collections of documents or materials with search functionality to facilitate easy access and exploration.

10. Curated Collection:

- Collections tailored to a particular theme or type of resource, such as art, postcards, or other materials, which may be part of a museum or other cultural institution.

11.and more….