

Recommended Formats Statement

Print Subscribe Share/Save Give Feedback

« [Back to Recommended Formats Statement](#)

[Main](#) | [Table of Contents](#) | [Introduction](#) | [Textual Works](#) | [Still Image Works](#) | [Moving Image Works](#) | [Audio Works](#) | [Musical Scores](#) | [Datasets](#) | [GIS, Geospatial and Non-GIS Cartographic](#) | [Design and 3D](#) | [Software and Video Games](#) | [Web Archives](#)

VI. Datasets

NOTE: See also [Geospatial and Cartographic](#)

The Library is aware that, in some cases, the provision of datasets and databases for current research uses (including support for the U.S. Congress) may depend upon native formats and associated software, while preservation and long-term access may depend upon data-migration via transport or export formats, with a concomitant risk of loss of precision and accuracy. Given the focus of this document is preservation and long-term access, the following format preferences favor those outcomes.

[i. Datasets](#)

i. Datasets		
	Preferred	Acceptable
A. Formats	<ol style="list-style-type: none"> Platform-independent, character-based formats are preferred over native or binary formats as long as data is complete, and retains full detail and precision. Preferred formats include well-developed, widely adopted, de facto marketplace standards, e.g. <ol style="list-style-type: none"> Formats using well known schemas with public validation tool available Line-oriented, e.g. TSV, CSV, fixed-width Platform-independent open formats, e.g. .db, .db3 Any proprietary format that is a de facto standard for a profession or supported by multiple tools (e.g. Excel .xls or .xlsx, Shapefile) Character Encoding, in descending order of preference: <ol style="list-style-type: none"> UTF-8, UTF-16 (with BOM), US-ASCII or ISO 8859-1 Other named encoding 	<p>In order of preference</p> <ol style="list-style-type: none"> Non-proprietary, publicly documented formats endorsed as standards by a professional community or government agency, e.g. CDF, HDF Text-based data formats with available schema
B. Related Materials	Consult the appropriate sections of this document to identify the preferred formats for supplementary material	
C. Delivery Method, in order of preference	<ol style="list-style-type: none"> Public download URLs Automated private download URLs with any necessary API keys or credentials Hard drive; CD-ROM; DVD-ROM 	
D. Metadata	<ol style="list-style-type: none"> Deposits should include all applicable metadata, data dictionaries, XML schemas, and technical specifications as appropriate. Discipline-specific metadata standards should be used whenever possible As supported by format: <ol style="list-style-type: none"> Title 	

	<ul style="list-style-type: none"> b. Creator c. Creation date d. Place of publication e. Publisher/producer/distributor f. Contact information g. A list of software used to produce, render or compress the data (if applicable) h. Character encoding <p>3. Include if available:</p> <ul style="list-style-type: none"> a. Language of work b. Other relevant identifiers (e.g., DOI, LCCN, canonical URL, etc.) c. Subject descriptors d. Abstracts e. Key or reference to each data field f. Checksums g. Permanent version specifiers h. Information about how the data was collected and any sampling or post-processing which as been applied i. Known copyright terms, especially for datasets which combine data from multiple sources <p>4. For datasets serving as part of a database: proprietary database package and version</p>	
<p>E. Technological Measures</p>	<p>Files must contain no measures (such as digital rights management technologies or encryption) that control access to or prevent use of the digital work.</p> <p>Files in formats which support linking or embedding external resources (e.g. XML, JSON, Excel .xls or .xlsx) should be self-contained to remain useful in the event of external service changes.</p> <p>Files in formats which support executable code (e.g. Excel) do not contain executable code.</p>	<p>Files in formats which support executable code do not depend on embedded programs for purposes other than display (e.g. search, filtering, etc.); the raw data is available without executing code.</p>

[⬅ Back to Top](#)

ii. Databases

ii. Databases		
	Preferred	Acceptable
A. Preservation	Complete set of the content contained within the database	
B. Access, in order of preference	<ul style="list-style-type: none"> 1. Publisher web interface with: <ul style="list-style-type: none"> a. Comprehensive and user-friendly search and discovery b. Counter compliant usage statistics 2. Delivered preservation content 	Documented API

[⬅ Back to Top](#)