Format Standards:

What Do I Need To Know?

Overview for Today:

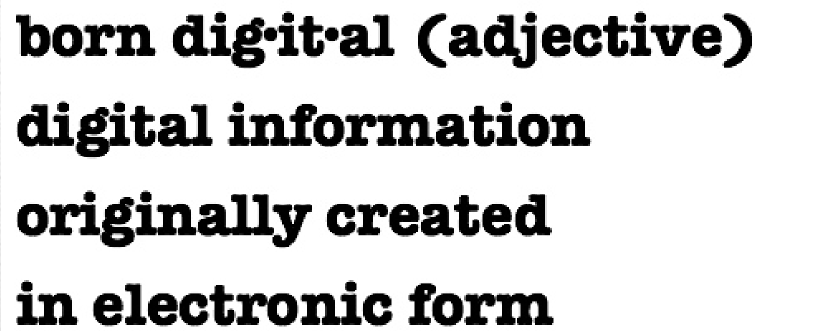
1. What are Formats and Why Should You Care?

2. Formats and Preservation Strategies for Textual Records, Structured Records, and Email.

3. Brief Overview of Strategies and Available Resources.

Understanding Formats

BORN-DIGITAL



born dig·it·al (adjective)
digital information
originally created
in electronic form

DIGITIZED

FORMAT DECISIONS AND ISSUES APPLY TO BOTH!

MEDIA

CONTENT

CSV

TXT

Adobe

Understanding Formats

File formats are the “rules that specify how the bytes that make up [a given] file are organized, interpreted, and rendered.”

- Ciran B. Trace

Understanding Formats

The Biggest Risk with Formats?

Understanding Formats

Ask questions about format choices now…

…to help insure future access.

Understanding Formats

QUESTION YOUR FORMATS

WHO OWNS?

WHO SUPPORTS?

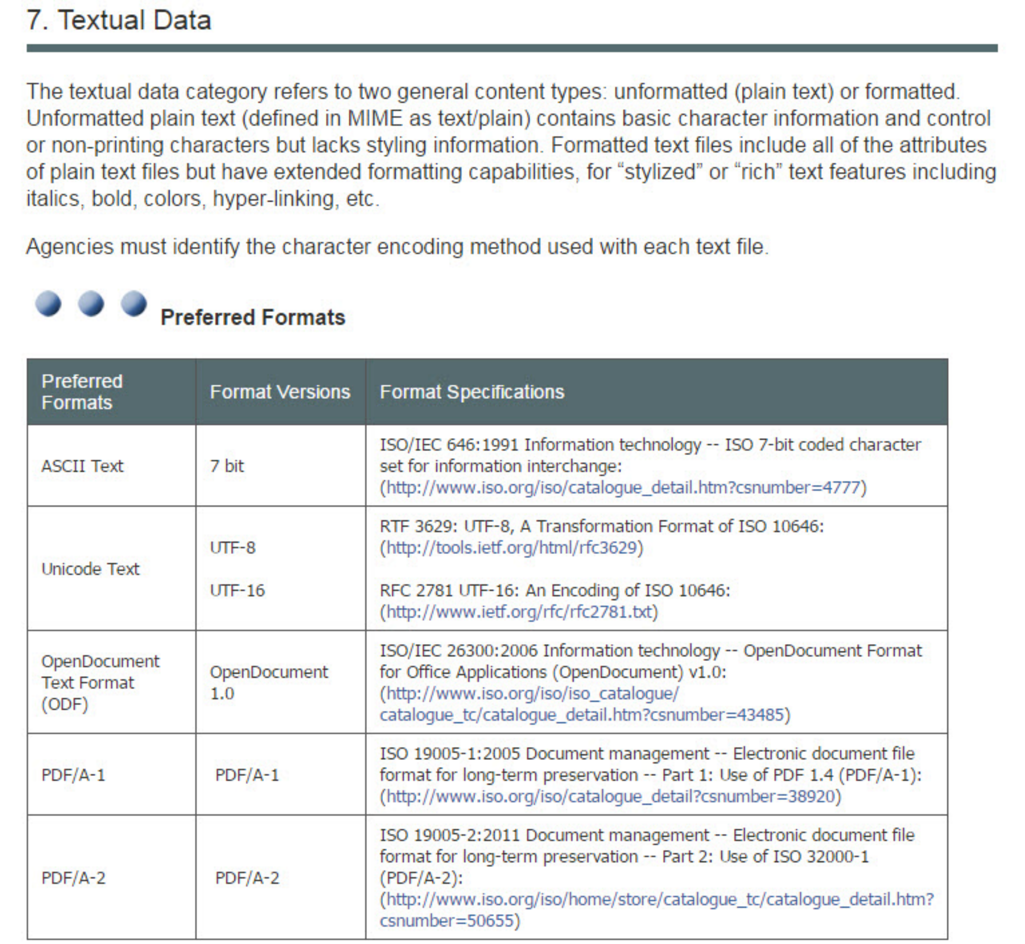
WHO (OR WHAT) GOVERNS?

Understanding Formats

Textual Records

NATIONAL ARCHIVES

- Refers to born digital textual data.
- Digitized text is a concern for image formats.
- Generally two types:
  - Formatted
  - Unformatted



7. Textual Data

The textual data category refers to two general content types: unformatted (plain text) or formatted. Unformatted plain text (defined in MIME as text/plain) contains basic character information and control or non-printing characters but lacks styling information. Formatted text files include all of the attributes of plain text files but have extended formatting capabilities, for “stylized” or “rich” text features including italics, bold, colors, hyper-linking, etc.

Agencies must identify the character encoding method used with each text file.

● ● ● **Preferred Formats**

Preferred Formats	Format Versions	Format Specifications
ASCII Text	7 bit	ISO/IEC 646:1991 Information technology -- ISO 7-bit coded character set for information interchange: (http://www.iso.org/iso/catalogue_detail.htm?csnumber=4777)
Unicode Text	UTF-8 UTF-16	RTF 3629: UTF-8, A Transformation Format of ISO 10646: (http://tools.ietf.org/html/rfc3629) RFC 2781 UTF-16: An Encoding of ISO 10646: (http://www.ietf.org/rfc/rfc2781.txt)
OpenDocument Text Format (ODF)	OpenDocument 1.0	ISO/IEC 26300:2006 Information technology -- OpenDocument Format for Office Applications (OpenDocument) v1.0: (http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=43485)
PDF/A-1	PDF/A-1	ISO 19005-1:2005 Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1): (http://www.iso.org/iso/catalogue_detail?csnumber=38920)
PDF/A-2	PDF/A-2	ISO 19005-2:2011 Document management -- Electronic document file format for long-term preservation -- Part 2: Use of ISO 32000-1 (PDF/A-2): (http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=50655)

Textual Records

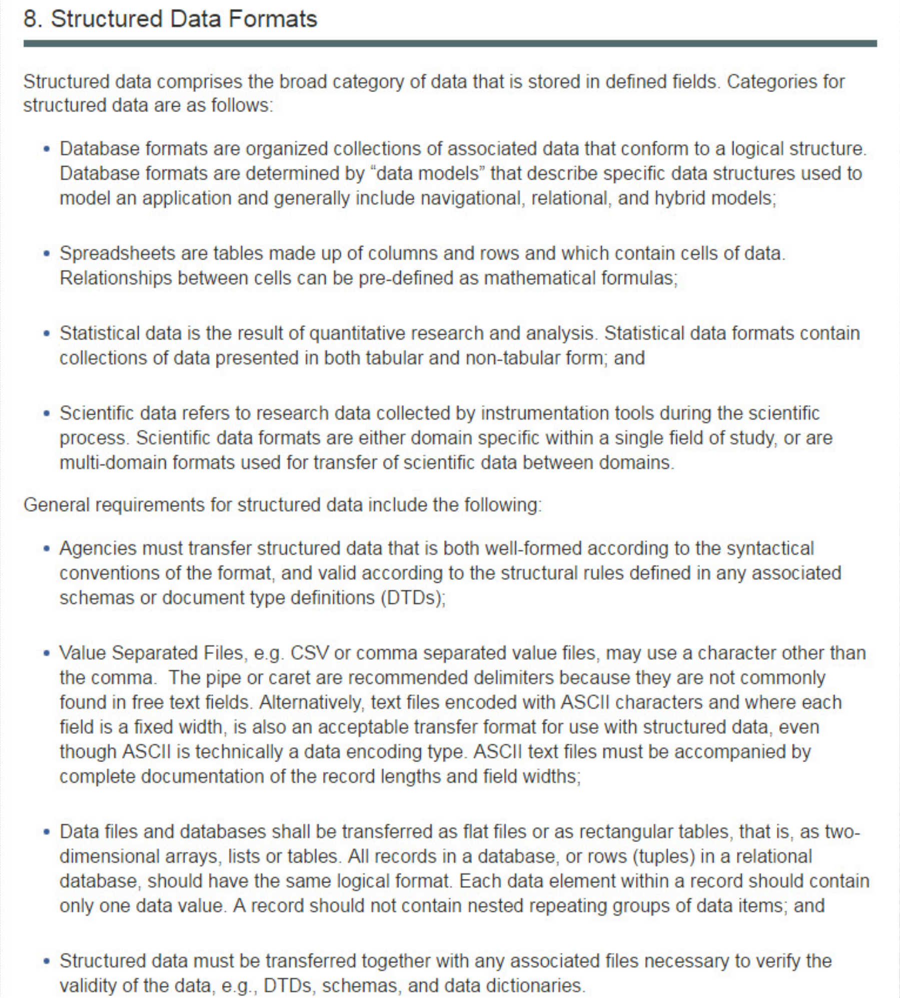
COMMON FORMATS

Structured Records

NATIONAL ARCHIVES

- Concerns data that is organized and stored in defined fields. This can include:
  - Databases
  - Spreadsheets
  - Statistical Data
  - Scientific data

- Typically requires transfer of other associated files and metadata (such as schemas or data dictionaries) that help make the structured data valid and comprehensible.



8. Structured Data Formats

Structured data comprises the broad category of data that is stored in defined fields. Categories for structured data are as follows:

- Database formats are organized collections of associated data that conform to a logical structure. Database formats are determined by "data models" that describe specific data structures used to model an application and generally include navigational, relational, and hybrid models;
- Spreadsheets are tables made up of columns and rows and which contain cells of data. Relationships between cells can be pre-defined as mathematical formulas;
- Statistical data is the result of quantitative research and analysis. Statistical data formats contain collections of data presented in both tabular and non-tabular form; and
- Scientific data refers to research data collected by instrumentation tools during the scientific process. Scientific data formats are either domain specific within a single field of study, or are multi-domain formats used for transfer of scientific data between domains.

General requirements for structured data include the following:

- Agencies must transfer structured data that is both well-formed according to the syntactical conventions of the format, and valid according to the structural rules defined in any associated schemas or document type definitions (DTDs);
- Value Separated Files, e.g. CSV or comma separated value files, may use a character other than the comma. The pipe or caret are recommended delimiters because they are not commonly found in free text fields. Alternatively, text files encoded with ASCII characters and where each field is a fixed width, is also an acceptable transfer format for use with structured data, even though ASCII is technically a data encoding type. ASCII text files must be accompanied by complete documentation of the record lengths and field widths;
- Data files and databases shall be transferred as flat files or as rectangular tables, that is, as two-dimensional arrays, lists or tables. All records in a database, or rows (tuples) in a relational database, should have the same logical format. Each data element within a record should contain only one data value. A record should not contain nested repeating groups of data items; and
- Structured data must be transferred together with any associated files necessary to verify the validity of the data, e.g., DTDs, schemas, and data dictionaries.

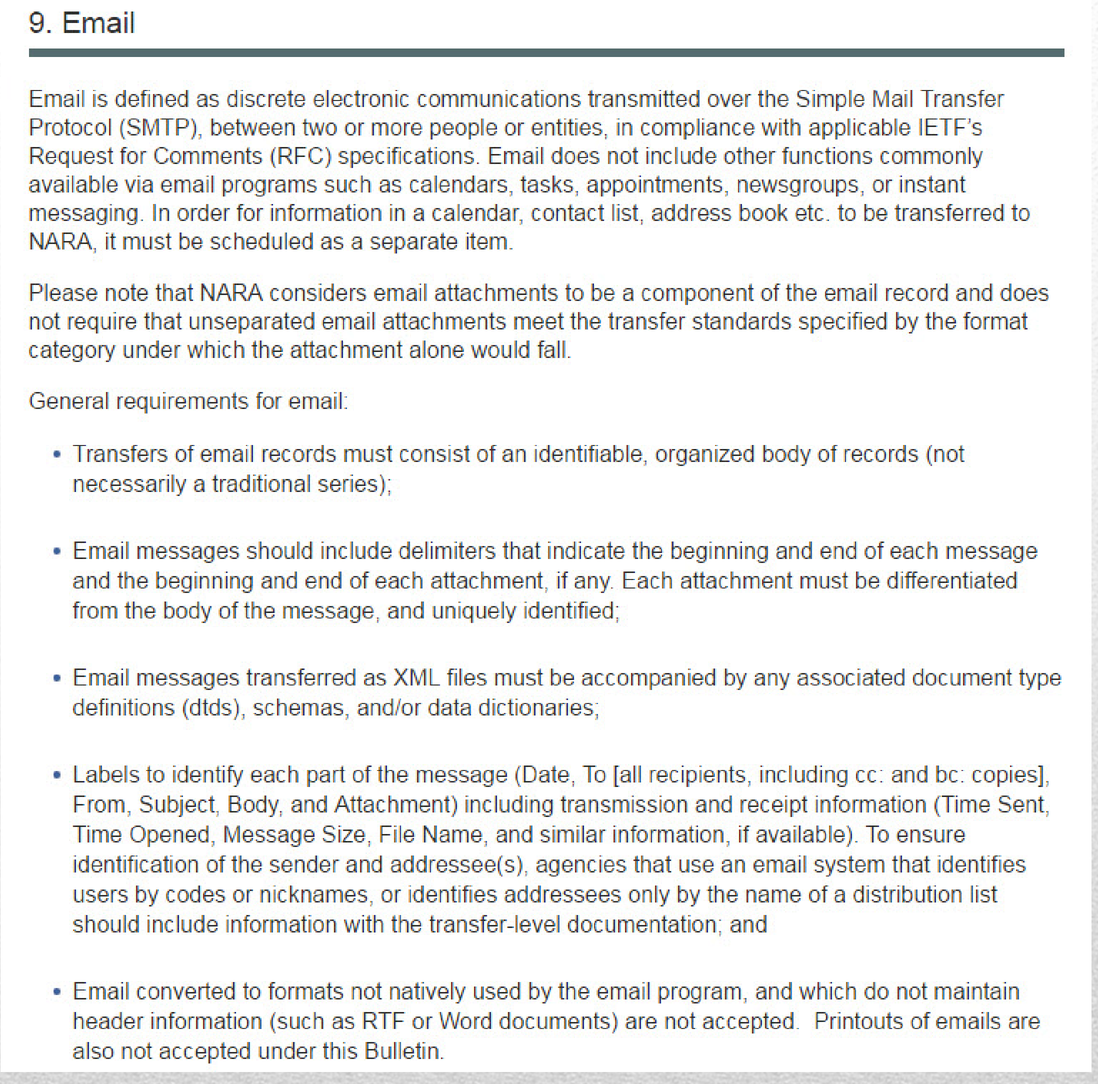
COMMON FORMATS

Structured Records

Email

NATIONAL ARCHIVES

- Consider the traditional functions email has come to replace.
- Critical that email is preserved and transferred according to set retention schedules and/or collection policies.
- Email is unique in how it can seem to transition between textual record and structured record.



9. Email

Email is defined as discrete electronic communications transmitted over the Simple Mail Transfer Protocol (SMTP), between two or more people or entities, in compliance with applicable IETF's Request for Comments (RFC) specifications. Email does not include other functions commonly available via email programs such as calendars, tasks, appointments, newsgroups, or instant messaging. In order for information in a calendar, contact list, address book etc. to be transferred to NARA, it must be scheduled as a separate item.

Please note that NARA considers email attachments to be a component of the email record and does not require that unseparated email attachments meet the transfer standards specified by the format category under which the attachment alone would fall.

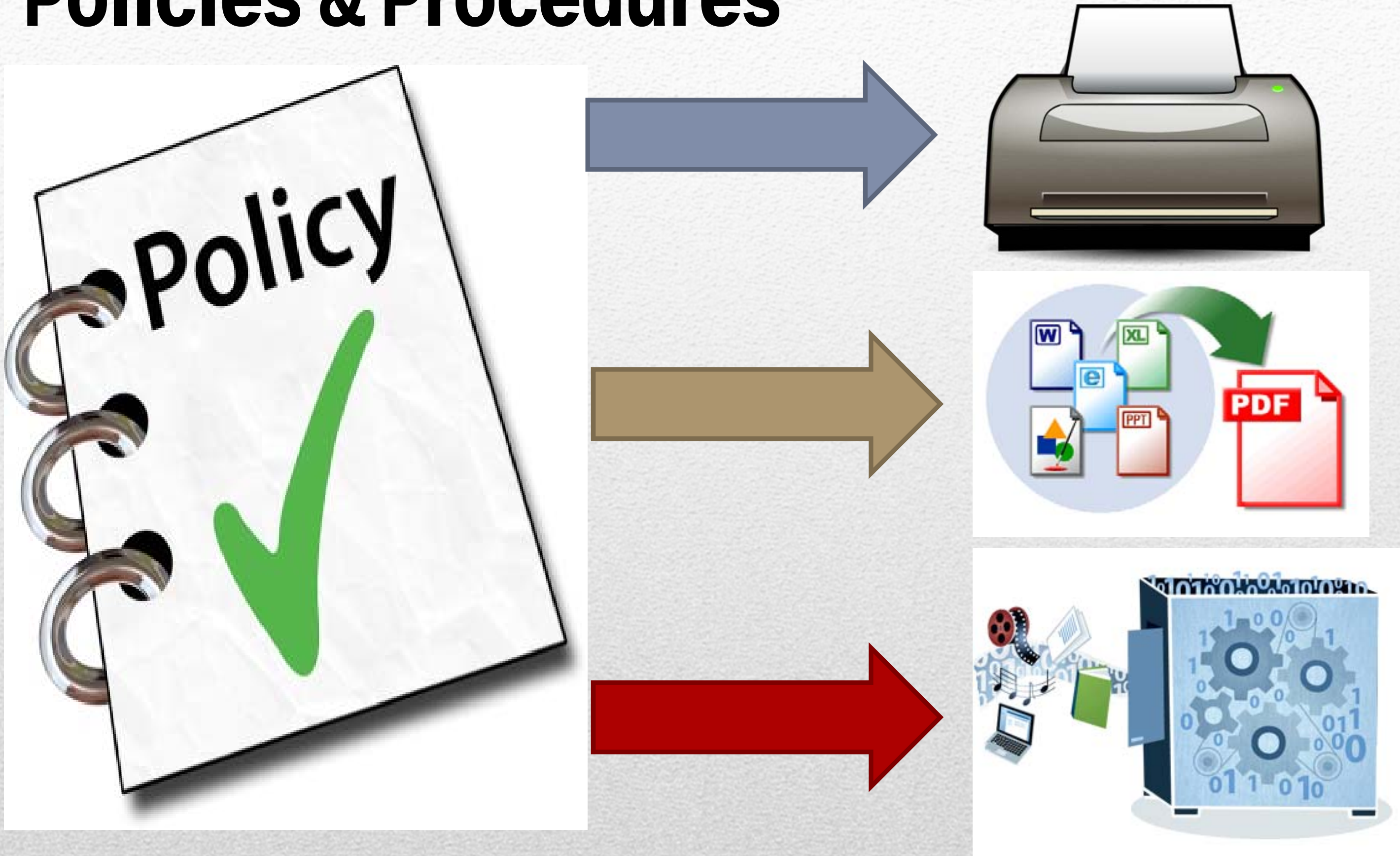
General requirements for email:

- Transfers of email records must consist of an identifiable, organized body of records (not necessarily a traditional series);
- Email messages should include delimiters that indicate the beginning and end of each message and the beginning and end of each attachment, if any. Each attachment must be differentiated from the body of the message, and uniquely identified;
- Email messages transferred as XML files must be accompanied by any associated document type definitions (dtds), schemas, and/or data dictionaries;
- Labels to identify each part of the message (Date, To [all recipients, including cc: and bc: copies], From, Subject, Body, and Attachment) including transmission and receipt information (Time Sent, Time Opened, Message Size, File Name, and similar information, if available). To ensure identification of the sender and addressee(s), agencies that use an email system that identifies users by codes or nicknames, or identifies addressees only by the name of a distribution list should include information with the transfer-level documentation; and
- Email converted to formats not natively used by the email program, and which do not maintain header information (such as RTF or Word documents) are not accepted. Printouts of emails are also not accepted under this Bulletin.

COMMON FORMATS

Email

Policies & Procedures

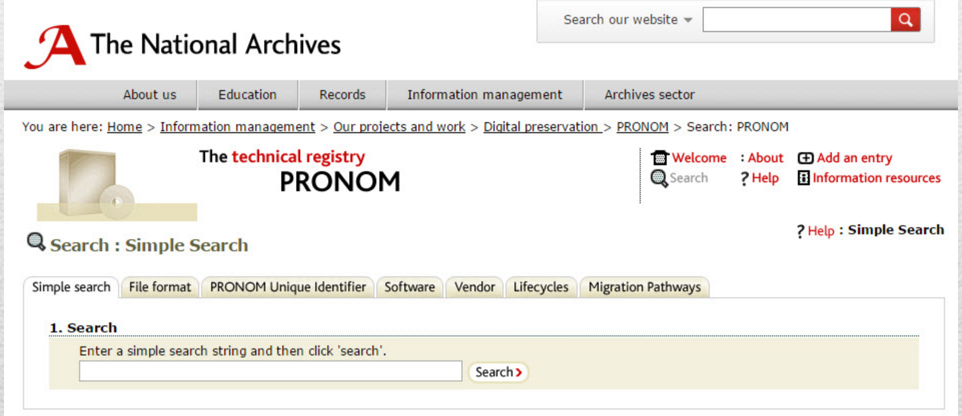


Strategies and Resources

Validation Tools

JHOVE *JSTOR/Harvard Object Validation Environment*

http://jhove.sourceforge.net/



You are here: Home > Information management > Our projects and work > Digital preservation > PRONOM > Search: PRONOM

The technical registry
PRONOM

Search : Simple Search

? Help : Simple Search

1. Search

Enter a simple search string and then click 'search'.

Home / Browse / Development / Data Formats / DROID

DROID

Brought to you by: davidunderdown, djclipsham

https://www.nationalarchives.gov.uk/PRONOM/Default.aspx

Strategies and Resources

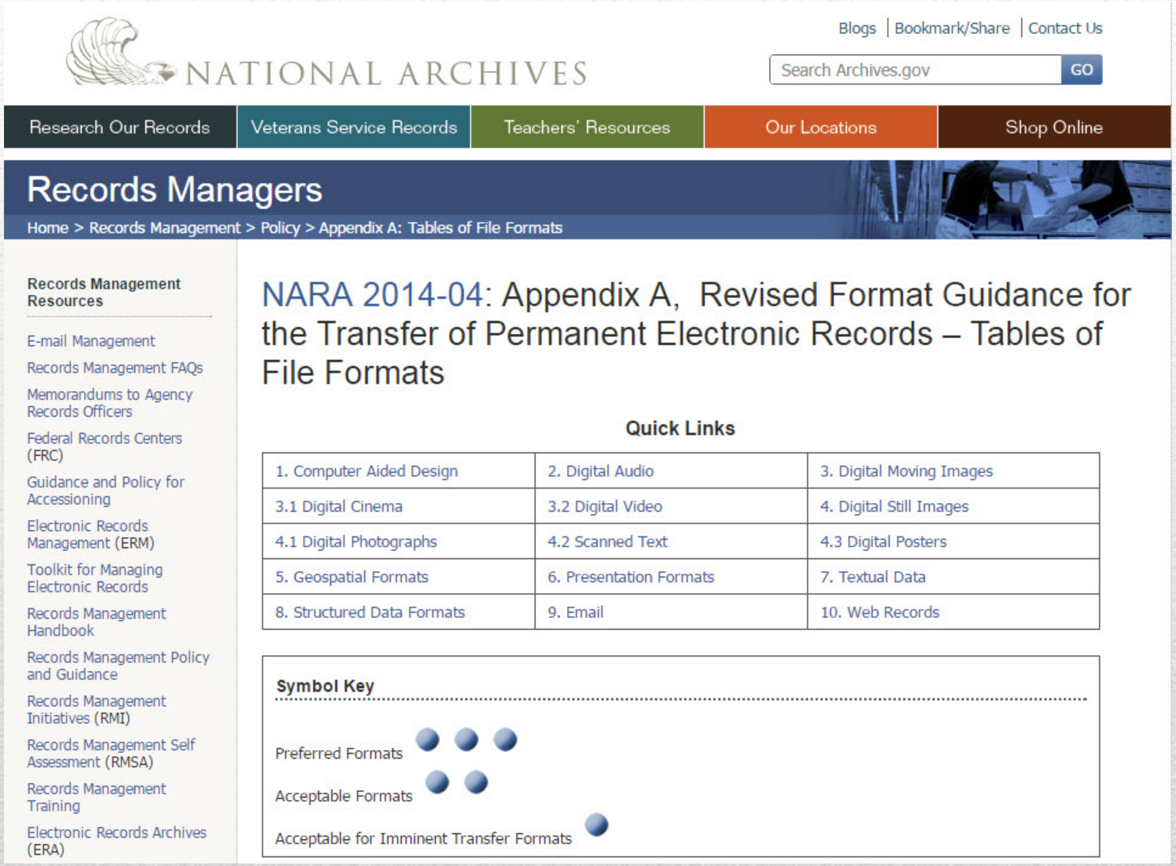
Open Source Software

https://www.openoffice.org/

https://sourceforge.net/projects/pdfcreator/

Strategies and Resources

Professional Guidance


Blogs | Bookmark/Share | Contact Us
NATIONAL ARCHIVES
Search Archives.gov
GO
Research Our Records
Veterans Service Records
Teachers' Resources
Our Locations
Shop Online
Records Managers
Home > Records Management > Policy > Appendix A: Tables of File Formats
Records Management Resources
E-mail Management
Records Management FAQs
Memorandums to Agency Records Officers
Federal Records Centers (FRC)
Guidance and Policy for Accessioning
Electronic Records Management (ERM)
Toolkit for Managing Electronic Records
Records Management Handbook
Records Management Policy and Guidance
Records Management Initiatives (RMI)
Records Management Self Assessment (RMSA)
Records Management Training
Electronic Records Archives (ERA)

NARA 2014-04: Appendix A, Revised Format Guidance for the Transfer of Permanent Electronic Records – Tables of File Formats

Quick Links

1. Computer Aided Design	2. Digital Audio	3. Digital Moving Images
3.1 Digital Cinema	3.2 Digital Video	4. Digital Still Images
4.1 Digital Photographs	4.2 Scanned Text	4.3 Digital Posters
5. Geospatial Formats	6. Presentation Formats	7. Textual Data
8. Structured Data Formats	9. Email	10. Web Records

Symbol Key

Preferred Formats	● ● ●
Acceptable Formats	● ●
Acceptable for Imminent Transfer Formats	●

NARA 2014-04: Appendix A

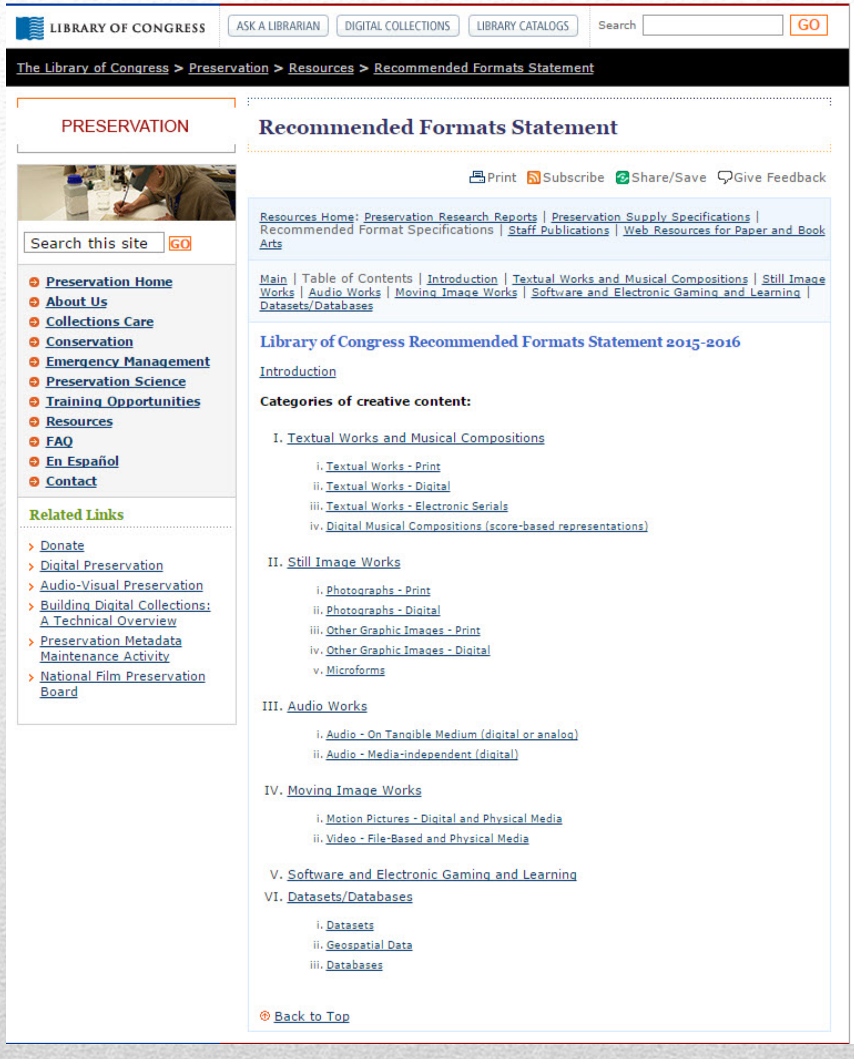
- Provides format standards for a variety of digital objects encountered in an archival digital preservation environment.
- Tiered categories for formats that provides greater flexibility.
- https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html#presentationformats

Strategies and Resources

Professional Guidance

LoC Recommended Formats Statement:

- "...purpose is to inform the creative and library communities on best practices for ensuring the preservation of, and long-term access to, the creative output of the nation and the world."
- Provides format recommendations and standards for both analog and digital materials.
- https://www.loc.gov/preservation/resources/rfs/


LIBRARY OF CONGRESS
ASK A LIBRARIAN
DIGITAL COLLECTIONS
LIBRARY CATALOGS
Search
GO
The Library of Congress > Preservation > Resources > Recommended Formats Statement
PRESERVATION
Search this site
GO
Preservation Home
About Us
Collections Care
Conservation
Emergency Management
Preservation Science
Training Opportunities
Resources
FAQ
En Español
Contact
Related Links
Donate
Digital Preservation
Audio-Visual Preservation
Building Digital Collections: A Technical Overview
Preservation Metadata Maintenance Activity
National Film Preservation Board
Print
Subscribe
Share/Save
Give Feedback
Resources Home: Preservation Research Reports | Preservation Supply Specifications | Recommended Format Specifications | Staff Publications | Web Resources for Paper and Book Arts
Main | Table of Contents | Introduction | Textual Works and Musical Compositions | Still Image Works | Audio Works | Moving Image Works | Software and Electronic Gaming and Learning | Datasets/Databases
Back to Top

Recommended Formats Statement

Library of Congress Recommended Formats Statement 2015-2016

Introduction

Categories of creative content:

I. Textual Works and Musical Compositions
  i. Textual Works - Print
  ii. Textual Works - Digital
  iii. Textual Works - Electronic Serials
  iv. Digital Musical Compositions (score-based representations)

II. Still Image Works
  i. Photographs - Print
  ii. Photographs - Digital
  iii. Other Graphic Images - Print
  iv. Other Graphic Images - Digital
  v. Microforms

III. Audio Works
  i. Audio - On Tangible Medium (digital or analog)
  ii. Audio - Media-independent (digital)

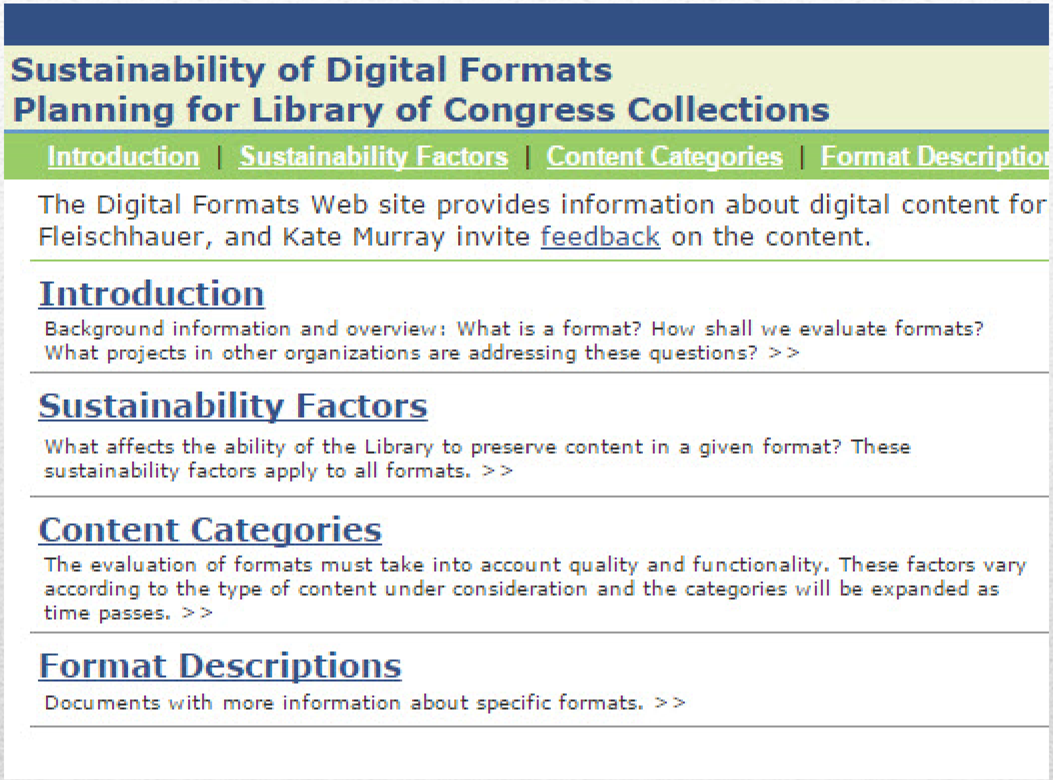
IV. Moving Image Works
  i. Motion Pictures - Digital and Physical Media
  ii. Video - File-Based and Physical Media

V. Software and Electronic Gaming and Learning

VI. Datasets/Databases
  i. Datasets
  ii. Geospatial Data
  iii. Databases

Strategies and Resources

Professional Guidance



Sustainability of Digital Formats
Planning for Library of Congress Collections

Introduction | Sustainability Factors | Content Categories | Format Descriptio

The Digital Formats Web site provides information about digital content for
Fleischhauer, and Kate Murray invite feedback on the content.

Introduction

Background information and overview: What is a format? How shall we evaluate formats? What projects in other organizations are addressing these questions? >>

Sustainability Factors

What affects the ability of the Library to preserve content in a given format? These sustainability factors apply to all formats. >>

Content Categories

The evaluation of formats must take into account quality and functionality. These factors vary according to the type of content under consideration and the categories will be expanded as time passes. >>

Format Descriptions

Documents with more information about specific formats. >>

LoC Sustainability of Digital Formats

- Useful tool for thinking about long-term format sustainability, and building corresponding institutional format policies.
- Discusses sustainability factors, content categories, and provides descriptions for individual content formats.
- http://www.digitalpreservation.gov/formats/

Strategies and Resources

James Kichas

jkichas@utah.gov

801-531-3844