NORTH CAROLINA DEPARTMENT OF CULTURAL RESOURCES
WWW.NCCULTURE.COM

# FILE FORMAT GUIDELINES FOR MANAGEMENT AND LONG-TERM RETENTION OF ELECTRONIC RECORDS

9/10/2012

State Archives of North Carolina

## Table of Contents

# File Format Guidelines for Management and Long-Term Retention of Electronic records

STATE ARCHIVES OF NORTH CAROLINA

## 1. GUIDELINES AND RECOMMENDATIONS

The following table represents the digital formats that the State Archives of North Carolina (State Archives) recommends for in-house preservation and long-term records retention. For electronic records, long-term retention is considered any period 3 - 5 years or longer. The State Archives recommends that any state or local agency record series for which the required retention period is five years or longer be maintained in the following formats. The record types included in this document are not exhaustive. State and local agencies producing specialized records may find that certain types of records are not covered by this document. Please contact the Electronic Records Branch to discuss potential preservation strategies for such media.

These guidelines classify formats into three categories:

**Recommended for long-term retention:** File formats that meet the minimum requirements for long-term retention, including documentation, wide adoption, transparency, self-containment, and use within the archival community. In most cases, these are the formats the State Archives itself uses to preserve electronic records.

**Acceptable for long-term retention**: File formats that do not meet the minimum requirements for long-term retention, but which come near to meeting the requirements and, for practical reasons, may be appropriate for long-term retention at some agencies. These formats are more likely to require frequent review and maintenance than formats recommended for long-term retention.

**Not recommended for long-term retention**: File formats that are not appropriate for long-term retention. Files saved in these formats should not be relied on to last more than five years. Electronic records whose retention periods are over five years should not be stored in these formats.

| Type of record | Recommended for long-term retention | Acceptable for long-term retention | Not recommended for long-term retention |
|---|---|---|---|
| **Word Processing documents** | PDF/A-1a (.pdf) (ISO 19005-1 compliant PDF/A)<br>OpenDocument Text (.odt) | PDF/A-1b (.pdf) (ISO 19005-1 minimally compliant PDF/A)<br>Microsoft® Word Document (.doc)<br>Microsoft® Open XML Document (.docx)<br>Rich Text Format (.rtf) | Corel® WordPerfect® (.wpd)<br>Lotus® WordPro (.lwp)<br>PDF (.pdf) |
| **Plain text documents** | Plain Text (.txt) *US-ASCII or UTF-8 encoding*<br>Comma-separated file (.csv) *US-ASCII or UTF-8 encoding*<br>Tab-delimited file (.txt) *US-ASCII or UTF-8 encoding* | Other delimited text files (space-delimited, colon-delimited, etc.) *where the delimiting character is not present in the data* | |
| **Structural markup text documents** | SGML *with DTD/Schema*<br>XML (.xml) *with DTD/Schema* | | XML without DTD/Schema<br>SGML without DTD/Schema |
| **Spreadsheets** | OpenDocument Spreadsheet (.ods)<br>Comma-separated file (.csv)<br>Tab-delimited file (.txt)<br>PDF/A-1a (.pdf) (ISO 19005-1 compliant PDF/A) | Microsoft® Excel® Spreadsheet (.xls)<br>Microsoft® Excel® Open XML Spreadsheet (.xlsx)<br>Other delimited text files (space-delimited, colon-delimited, etc.) *where the delimiting character is not present in the data* | |
| **Audio** | Broadcast WAVE Format LPCM (.wav)<br>WAVE Format LPCM (.wav) | AIFF (uncompressed) (.aif, .aiff)<br>Standard MIDI (.mid, .midi)<br>Windows® Media Audio WMA (. wma)<br>MPEG3  (.mp3)<br>MP4 AAC (.m4a) | Audio CD (Compact Disc Digital Audio system, CDDA, CD-DA)<br>DVD-Audio<br>QuickTime® MP4 AAC Protected (.m4p, .m4b)<br>QuickTime® MP3, iTunes (.mp3)<br>RealAudio® (.rm, .ra)<br>Shorten® (.shn)<br>RIFF-RMID (.rmi)<br>Extended MIDI (.xmi)<br>Module Music Formats, Mods (.mod)<br>SUN Audio, uncompressed (.au)<br>Ogg FLAC (.ogg) |

| Type of record | Recommended for long-term retention | Acceptable for long-term retention | Not recommended for long-term retention |
|---|---|---|---|
| **Digital Video** | AVI, full frame (uncompressed), WAVE PCM audio (.avi) | AVI, containing H.264/MPEG-4 AVC (lossy)[1] (.avi)<br>MPEG-4, containing H.264/MPEG-4 AVC (lossy) (.mp4)<br>MPEG-2, containing H.262/MPEG-2 (lossy) (.mp2)<br>MOV, containing H.264/MPEG-4 AVC (lossy) (.mov)<br>ASF, containing WMV (lossy) (.wmv)<br>MXF, containing Motion JPG 2000[2] (lossless) (.mxf)<br>Ogg, containing Theora (lossy) (.ogg) | DVD-Video<br>VOB (VIDEO_TS, AUDIO_TS)<br>Blu-ray Disc™<br>HCAM®<br>Digital VHS (D-VHS)<br>DVCam® |
| **Raster Images** | TIFF (.tif, .tiff) *uncompressed*<br>JPG 2000 (.jp2) | JPEG (.jpg, .jpeg)<br>PNG (.png)<br>PDF/A-1a (.pdf) (ISO 19005-1 compliant PDF/A)<br>GIF (.gif) | RAW (.raw, various)<br>Adobe® Photoshop® (.psd)<br>Kodak PhotoCD<br>Encapsulated PostScript (.eps)<br>FlashPix™ (.fpx)<br>PDF (.pdf) |
| **Vector Images** *(*See below for geospatial vector sets.)* | Scalable Vector Graphics 1.1 (.svg)<br>AutoCAD® Drawing Interchange Format (.dxf)<br>PDF/A-1a (.pdf) (ISO 19005-1 compliant PDF/A) | AutoCAD® Drawing Format (.dwg) | Adobe® Illustrator (.ai)<br>Corel®Draw CDR (.cdr)<br>Micrografx Draw DRW (.dwr)<br>Windows® Metafile WMF (.wmf, .emf)<br>Standard for the Exchange of Product Model Data STEP (.stp)<br>Computer Graphics Metafile DXF (.dxf) |
| **Databases** | Software Independent Archiving of Relational Databases (SIARD)<br>Delimited Flat File (Plain Text) with DDL | Microsoft® Access® (.accdb)<br>Microsoft® Access® (.mdb)<br>dBase Format (.dbf) | |
| **Presentations** | OpenDocument Presentation (.odp)<br>PDF/A-1a (.pdf) (ISO 19005-1 compliant PDF/A) *for presentations without animation* | | Microsoft® PowerPoint Presentation (.ppt)<br>Microsoft® Open XML PowerPoint® Presentation (.pptx) |

[1] One of the H.264/MPEG-4 AVC profiles is sometimes described as lossless: MPEG-4 AVC High 4:4:4 Profile. For more information, see the Notes section of "MPEG-4, Advanced Video Coding, High 4:4:4 Profile," *Sustainability of Digital Formats: Planning for Library of Congress Collections*, http://digitalpreservation.gov/formats/fdd/fdd000218.shtml (accessed 5/16/2012).

[2] See "MXF File, OP1a, Lossless JPEG 2000 in Generic Container," *Sustainability of Digital Formats: Planning for Library of Congress Collections*, http://digitalpreservation.gov/formats/fdd/fdd000206.shtml (accessed 5/16/2012) and "Motion JPEG 2000 jp2 File Format," *Sustainability of Digital Formats: Planning for Library of Congress Collections*, http://digitalpreservation.gov/formats/fdd/fdd000127.shtml (accessed 5/16/2012).

| Type of record | |
|---|---|
| **Email** | * *See section [2.11 Email](#).* |
| **Websites / Social Media** | * *See section [2.12 Webpages](#).* |
| **Geospatial Vector Data** | * *See section [2.13 Geospatial Vector Datasets](#).* |

# 2. DESCRIPTION OF FORMATS RECOMMENDED FOR LONG-TERM RETENTION

This section describes in further details the formats listed in column one of 1. Guidelines and Recommendations. These file formats meet the minimum requirements for long-term retention, including documentation, wide adoption, transparency, self-containment, and use within the archival community. In most cases, these are the formats the State Archives itself uses to preserve electronic records.

File formats are organized according to type of record, as presented in 1. Guidelines and Recommendations:

- Word Processing Documents
- Plain Text Documents
- Structural Markup Text Documents
- Spreadsheets
- Audio
- Digital Video
- Raster Images
- Vector Images
- Databases
- Presentations
- Email
- Websites/Social Media

## 2.1 Word Processing Documents

This category includes texts created in word processing applications like Microsoft® Word and OpenOffice. Unlike plain text files, these documents combine plain text with formatting and styling—including fonts, headings, lists, highlights, notes, and embedded tables and images.

NOTE: Although some word processing files, such as .docx files, are XML-based, for the purposes of these guidelines, these files have been included in the “word processing documents” category and distinguished from “structural markup text documents” due to differences in function and editing software (see 2.3 Structural Markup Text Documents).

### 2.1.1 PDF/A-1a (.pdf) (ISO 19005-1 compliant PDF/A)

PDF/A-1a, also known as “ISO 19005-1 compliant PDF/A” is a type of PDF document designed to preserve PDF files for long-term retention. Traditional PDF files have a number of weaknesses that can cause the same file to appear or behave differently when opened on different computers. Compliant PDF/A files overcome these issues and ensure that the PDF file will appear the same everywhere it is opened. Documents produced in word processing software like Microsoft® Word or WordPerfect should be converted to compliant PDF/A files.

PDF-Archival (more commonly known as PDF/A) is an international standard developed by the Association for Information and Image Management International (AIIM International) to archive and preserve electronic documents in PDF form. The PDF/A format has been adopted as ISO standard 19005-1:2005 and is widely used by archival institutions, including the National Archives and Records Administration (NARA), Library and Archives Canada (LAC), and the Library of Congress (LOC). Version 1, PDF/A-1, is the current archival standard.[3] It imposes several restrictions on the standard PDF format in order to maximize files' device independence, self-containment, and self-documentation. Constraints include:

1. Audio and video content are forbidden.
2. Javascript and executable file launches are prohibited.
3. All fonts must be embedded and also must be legally embeddable for unlimited, universal rendering (not under copyright).
4. Colors must be defined according to a universally available, device-independent color model.
5. Encryption is disallowed.
6. Image transparency is disallowed.
7. Use of standards-based metadata and tagging is mandated. This tagging makes documents understandable to screen readers; without it, documents cannot be Section 508 compliant.[4]

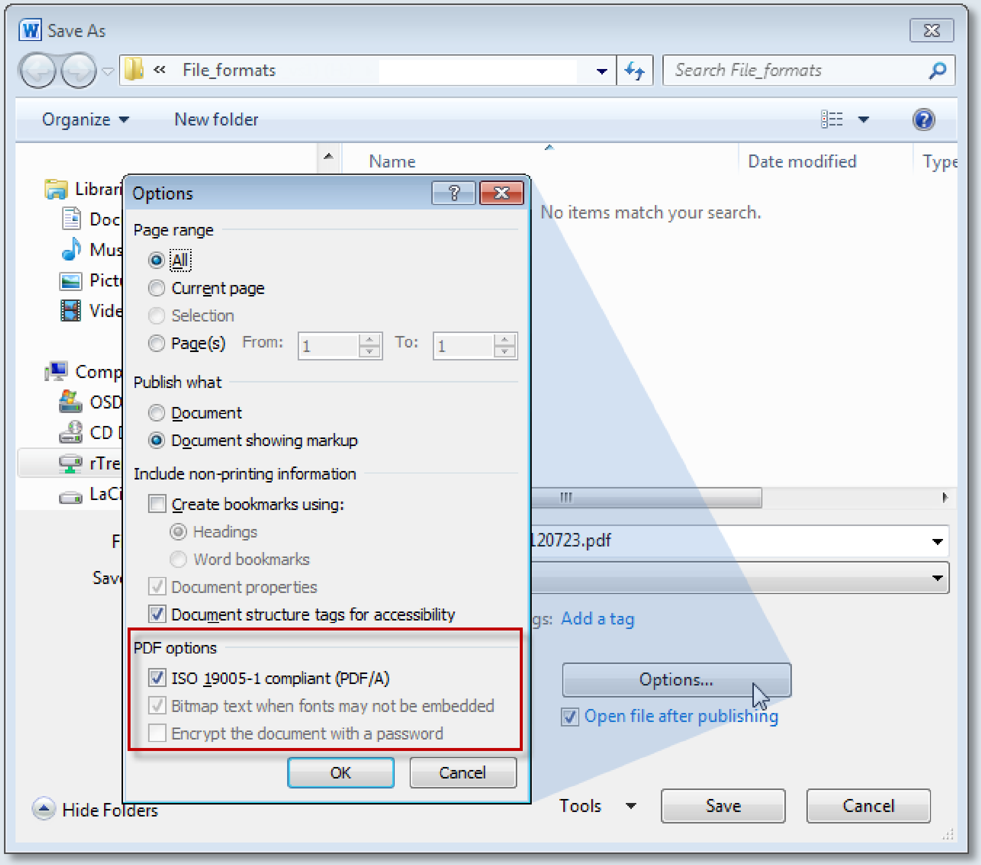
PDF/A-1 has two levels of compliance. The State Archives uses PDF/A-1a as a preservation standard. This level indicates "full compliance" with the restrictions listed above.

- **PDF/A-1a** — "full compliance" with the PDF/A standard. Typically, this is the default setting to which word processing software will save to PDF/A. Another way applications may describe PDF/A-1a is as "ISO 19005-1 compliant PDF/A."

- **PDF/A-1b** — "minimal compliance" with the PDF/A standard. PDF/A-1b ensures that the document will look the same in the future (preserves rendering), but it does not preserve the markup of the document.

[3] There is also a PDF/A version 2, or PDF/A-2, which was also adopted by ISO on June 20, 2011. As a new format, it has not been widely adopted in the archival community and it is still being investigated for archival and long-term preservation use.

[4] Library of Congress, "PDF/A-1, PDF for Long-term Preservation, Use of PDF 1.4," *Sustainability of Digital Formats: Planning for Library of Congress Collections*, http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml (accessed 02/17/2012); Daniel Noonan, Amy McCroy, and Elizabeth L. Black, "PDF/A: A Viable Addition to the Preservation Toolkit," *D-Lib Magazine*, 16.11/12 (November/December 2010), http://www.dlib.org/dlib/november10/noonan/11noonan.html (accessed 4/19/2012).



☑ Open file after publishing

Options

Page range

● All
○ Current page
○ Selection
○ Page(s) From: 1 To: 1

Publish what

○ Document
● Document showing markup

Include non-printing information

□ Create bookmarks using:
● Headings
○ Word bookmarks
☑ Document properties
☑ Document structure tags for accessibility

PDF options

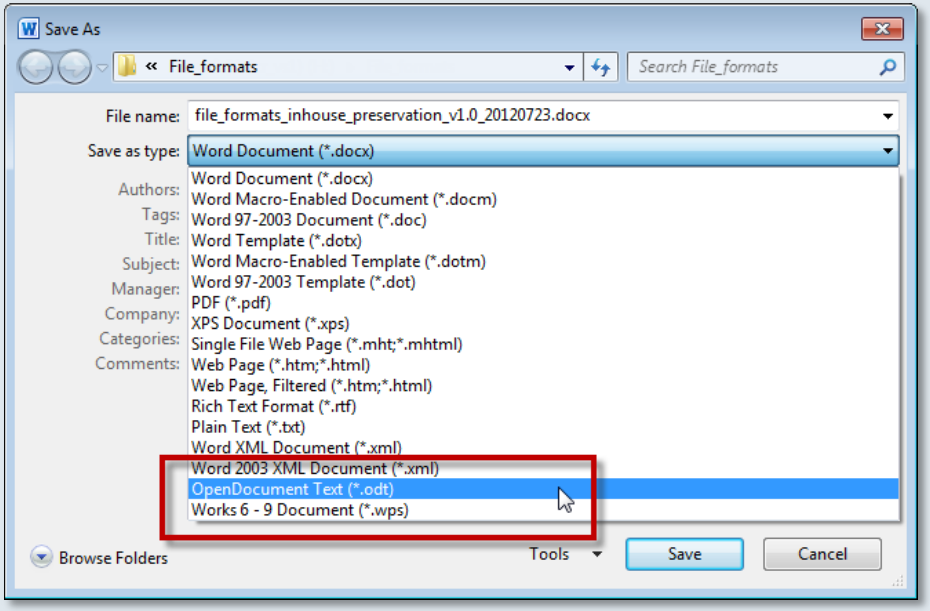
☑ ISO 19005-1 compliant (PDF/A)
☑ Bitmap text when fonts may not be embedded
□ Encrypt the document with a password

OK Cancel

**TIP: When saving a word processing document as a PDF, you should save the document as ISO 19005-1 compliant PDF/A file, also known as "PDF/A-1a." In the example above from Microsoft® Office 2010, this was accomplished by selecting PDF file format from the drop down menu, clicking the "Options…" button, and ticking the box labeled "ISO 19005-1 compliant (PDF/A)".**

### 2.1.2 OpenDocument Text (.odt)

OpenDocument Text is another preservation-quality format in which word processing document may be retained long-term. OpenDocument Text is similar in structure to the .docx format used by Microsoft® Office. OpenDocument Text is an open, non-proprietary format associated with many word processing applications, including OpenOffice. Most word processing applications can save and convert files to the OpenDocument Text format.[5]



File name: file_formats_inhouse_preservation_v1.0_20120723.docx

Save as type: Word Document (*.docx)

Save As
« File_formats
Search File_formats
Authors:
Tags:
Title:
Subject:
Manager:
Company:
Categories:
Comments:
Word Document (*.docx)
Word Macro-Enabled Document (*.docm)
Word 97-2003 Document (*.doc)
Word Template (*.dotx)
Word Macro-Enabled Template (*.dotm)
Word 97-2003 Template (*.dot)
PDF (*.pdf)
XPS Document (*.xps)
Single File Web Page (*.mht;*.mhtml)
Web Page (*.htm;*.html)
Web Page, Filtered (*.htm;*.html)
Rich Text Format (*.rtf)
Plain Text (*.txt)
Word XML Document (*.xml)
Word 2003 XML Document (*.xml)
OpenDocument Text (*.odt)
Works 6 - 9 Document (*.wps)
Browse Folders
Tools
Save
Cancel

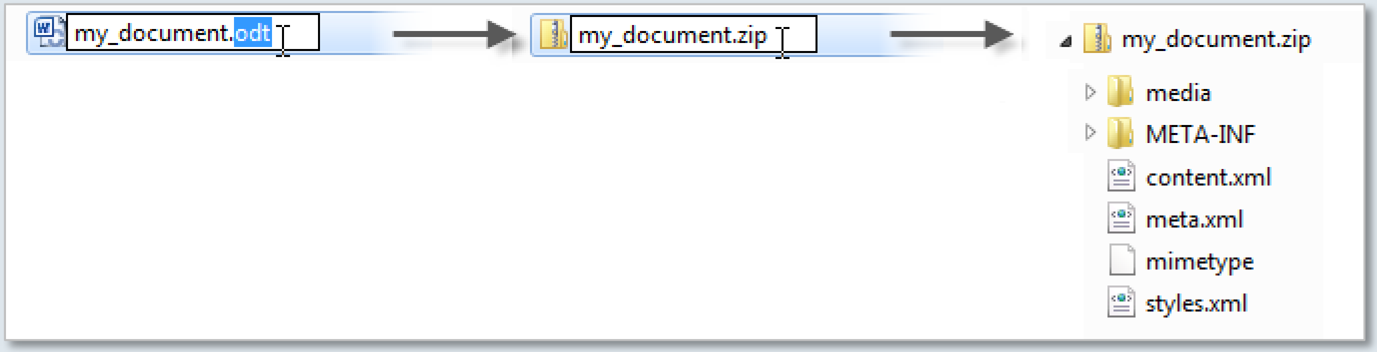
**Most word processing applications can save and convert files to the OpenDocument Text format, including Microsoft® Office 2010.**

[5] See also *OASIS Open Document Format for Office Applications (OpenDocument) TC,* https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office (accessed 9/7/2012).

OpenDocument Text is a sub-type of the OpenDocument Format (ODF), an open source file format for spreadsheets, charts, presentations, and word processing documents. Originally created by Sun Microsystems, the current standards were developed by the Organization for the Advancement of Structured Information Standards (OASIS) Open Document Format for Office Applications committee. The format is based on the XML format used by the OpenOffice.org office suite. The format is also published (in one of its version 1.0 manifestations) as ISO/IED international standard 26300:2006. See also 2.4.1 OpenDocument Spreadsheet (.ods) and 2.10.1 OpenDocument Presentation (.odp).

In almost all cases, an OpenDocument Text "file" with the .odt extension is actually a package of several files that have been compressed into a single ZIP file package that carries the .odt extension rather than the .zip extension. Within the zipped package are several separate files that represent the content of the document, its styling, metadata, settings, and a manifest of the zip package files. Although rare, an OpenDocument Text file can also be a single, flat XML file, in which case the associated file extension is usually .xml or .fodt.

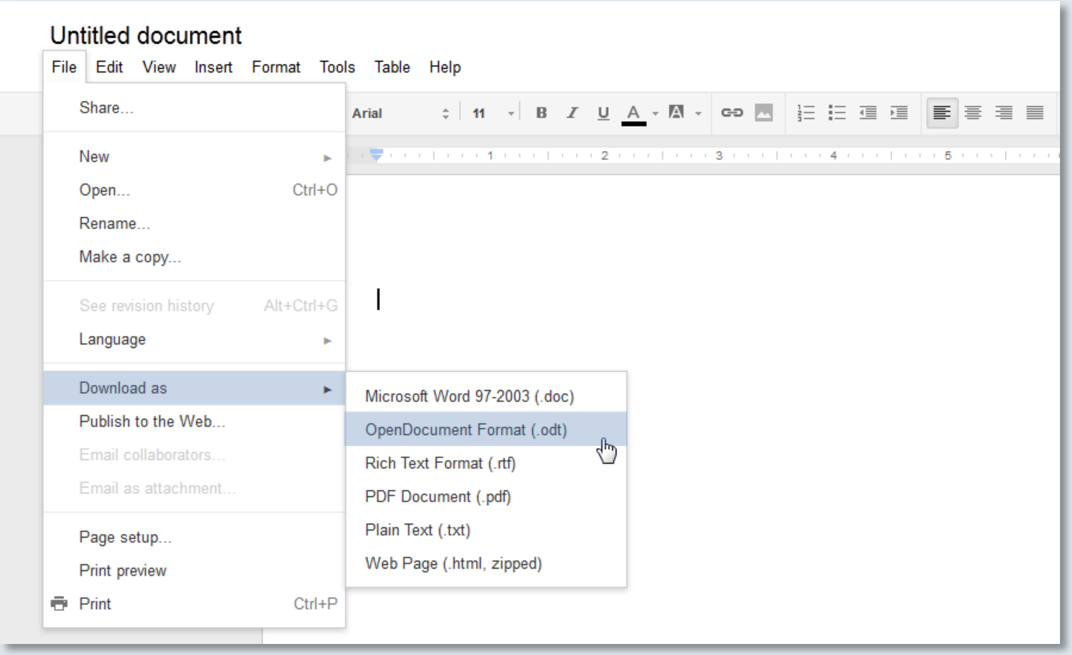
**TIP: You can see the internal structure of your OpenDocument Text (.odt) files if you change the extension from .odt to .zip, and then unzip the file. BE CAREFUL TO TRY THIS ONLY WITH TEST FILES, NOT PRESERVATION MASTER COPIES.**



### 2.1.3 *Special Note on Google Docs™*

Google Docs™ is a cloud-based document editing service offered by Google™. Word processing documents may be created on Google Docs™ and exported in various formats, including Microsoft® Word 97-2003 (.doc), OpenDocument Text (.odt), PDF (.pdf), zipped webpage (.zip), and others. The recommendations described in this document apply to all documents, regardless of whether they were created using

Google Docs™. The State Archives of North Carolina recommends that documents be exported from Google Docs™ as OpenDocument Format (.odt). Alternatively, documents can be exported as standard PDF files and then converted to PDF/A.



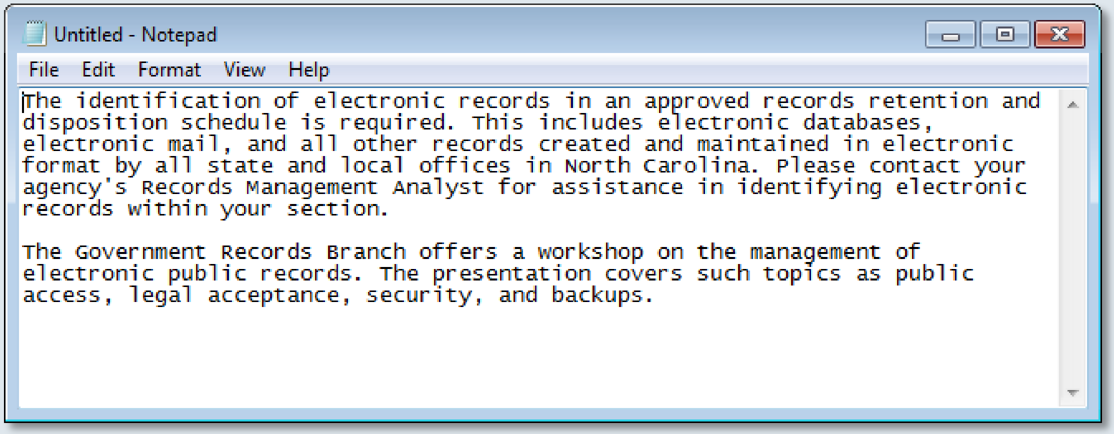
**Documents created in Google Docs™ should be exported for long-term preservation as OpenDocument Format (.odt). Alternatively, documents can be exported as standard PDF files and then converted to PDF/A.**

## 2.2 Plain Text Documents

Plain text files are those that contain US-ASCII or Unicode UTF-8 text without styling or structural markup. These are files commonly created with Notepad on Windows® operating systems, TextEdit on Mac® OS X® systems, and Vi text editor on Unix. Numerous other applicants are also used to create and edit these files. Technically speaking, while XML, HTML, XHTML, SGML, and many other documents are also plain text documents (typically Unicode UTF-8 encoded), these types of files utilize special markup languages to apply structural and styling rules to the documents' content. Because of their unique nature, such documents are classified for the purposes of these guidelines as "Structural markup text documents" (see 2.3 Structural Markup Text Documents).

### 2.2.1 Plain Text (.txt) *US-ASCII or UTF-8 encoding*

The data in plain text files is typically encoded in either US-ASCII or Unicode UTF-8 encodings. US-ASCII (American Standard Code for Information Interchange) defines 256 characters where each character is defined using an 8-bit byte. It is the most common encoding for English-language plain text documents. Unicode UTF-8 has a much broader set of characters, allowing for the use of non-Roman scripts (Arabic, Chinese, and Thai, for instance). Its first 128 characters are those used by US-ASCII, making Unicode UTF-8 backwards compatible with US-ASCII and making all US-ASCII text valid Unicode UTF-8 as well. Unicode UTF-8 has become the standard encoding for Web documents, including email.



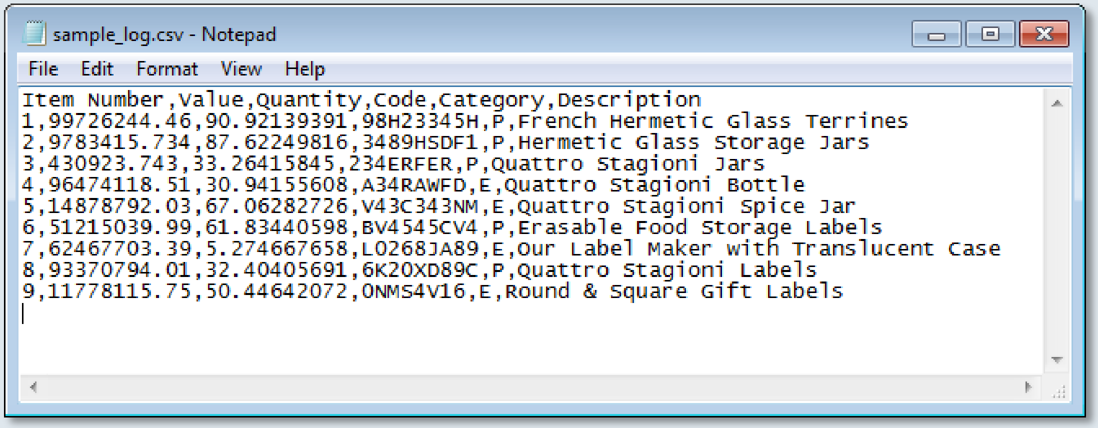
```
The identification of electronic records in an approved records retention and
disposition schedule is required. This includes electronic databases,
electronic mail, and all other records created and maintained in electronic
format by all state and local offices in North Carolina. Please contact your
agency's Records Management Analyst for assistance in identifying electronic
records within your section.

The Government Records Branch offers a workshop on the management of
electronic public records. The presentation covers such topics as public
access, legal acceptance, security, and backups.
```

**Example plain text file, opened in Notepad. See also comma-separated and tab-delimited files (below), which are types of plain text files designed especially to hold data.**

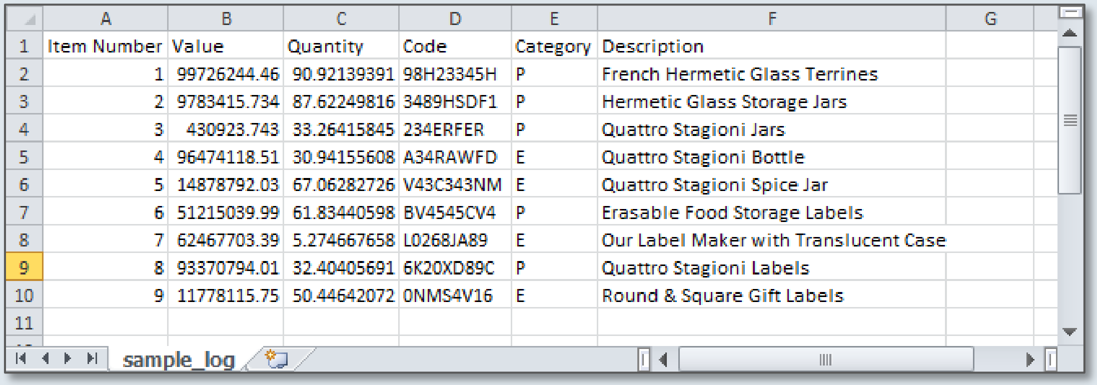
### 2.2.2 Comma-separated file (.csv) *US-ASCII or UTF-8 encoding*

Comma-separated files are plain text files that store tabular data. Like files with the .txt extension, they are usually encoded in either US-ASCII or Unicode UTF-8. They are distinguished by the fact that they contain values separated by commas and line breaks, so that spreadsheet and database applications (like Microsoft® Excel® and Access®) can easily open and interpret (or "parse") the data.



```
Item Number,Value,Quantity,Code,Category,Description
1,99726244.46,90.92139391,98H23345H,P,French Hermetic Glass Terrines
2,9783415.734,87.62249816,3489HSDF1,P,Hermetic Glass Storage Jars
3,430923.743,33.26415845,234ERFER,P,Quattro Stagioni Jars
4,96474118.51,30.94155608,A34RAWFD,E,Quattro Stagioni Bottle
5,14878792.03,67.06282726,V43C343NM,E,Quattro Stagioni Spice Jar
6,51215039.99,61.83440598,BV4545CV4,P,Erasable Food Storage Labels
7,62467703.39,5.274667658,L0268JA89,E,Our Label Maker with Translucent Case
8,93370794.01,32.40405691,6K20XD89C,P,Quattro Stagioni Labels
9,11778115.75,50.44642072,0NMS4V16,E,Round & Square Gift Labels
```

**Example comma-separated file (sample_log.csv) opened in Notepad. If this file were opened in a spreadsheet editor and re-saved as a tab-delimited file, it would appear as the file pictured in 2.2.3.**



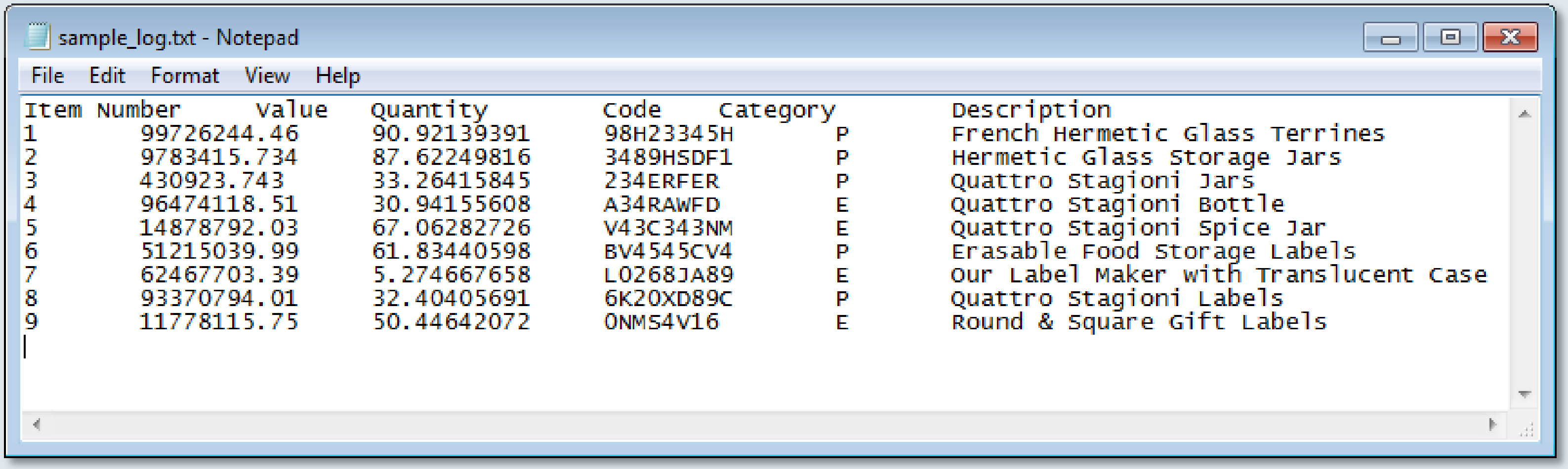
|  | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Item Number | Value | Quantity | Code | Category | Description |  |
| 2 | 1 | 99726244.46 | 90.92139391 | 98H23345H | P | French Hermetic Glass Terrines |  |
| 3 | 2 | 9783415.734 | 87.62249816 | 3489HSDF1 | P | Hermetic Glass Storage Jars |  |
| 4 | 3 | 430923.743 | 33.26415845 | 234ERFER | P | Quattro Stagioni Jars |  |
| 5 | 4 | 96474118.51 | 30.94155608 | A34RAWFD | E | Quattro Stagioni Bottle |  |
| 6 | 5 | 14878792.03 | 67.06282726 | V43C343NM | E | Quattro Stagioni Spice Jar |  |
| 7 | 6 | 51215039.99 | 61.83440598 | BV4545CV4 | P | Erasable Food Storage Labels |  |
| 8 | 7 | 62467703.39 | 5.274667658 | L0268JA89 | E | Our Label Maker with Translucent Case |  |
| 9 | 8 | 93370794.01 | 32.40405691 | 6K20XD89C | P | Quattro Stagioni Labels |  |
| 10 | 9 | 11778115.75 | 50.44642072 | 0NMS4V16 | E | Round & Square Gift Labels |  |
| 11 |  |  |  |  |  |  |  |

**The same comma-delimited file as above (sample_log.csv) opened in Microsoft® Excel®.**

### 2.2.3 Tab-delimited file (.txt) *US-ASCII or UTF-8 encoding*

Tab-delimited files are similar to comma separated files, the difference being that the values in one are separated by commas and in the other by tabs. Tab-delimited files carry the standard .txt extension.

As with the .txt files and comma-separated files described in 2.2.1 and 2.2.2, tab-delimited files should be encoded in either US-ASCII or Unicode UTF-8.



```
Item Number      Value    Quantity        Code    Category          Description
1       99726244.46     90.92139391     98H23345H       P       French Hermetic Glass Terrines
2       9783415.734     87.62249816     3489HSDF1       P       Hermetic Glass Storage Jars
3       430923.743      33.26415845     234ERFER        P       Quattro Stagioni Jars
4       96474118.51     30.94155608     A34RAWFD        E       Quattro Stagioni Bottle
5       14878792.03     67.06282726     V43C343NM       E       Quattro Stagioni Spice Jar
6       51215039.99     61.83440598     BV4545CV4       P       Erasable Food Storage Labels
7       62467703.39     5.274667658     L0268JA89       E       Our Label Maker with Translucent Case
8       93370794.01     32.40405691     6K20XD89C       P       Quattro Stagioni Labels
9       11778115.75     50.44642072     0NMS4V16        E       Round & Square Gift Labels
```

**Example tab-delimited file (sample_log.txt) opened in Notepad. If this file were opened in a spreadsheet editor and re-saved as a comma-separated file, it would appear as the files pictured in 2.2.2.**

## 2.3 Structural Markup Text Documents

Structural markup text documents, including XML and SGML, have been distinguished from plain text and word processing documents because of the unique functions they serve and the preservation standards they require. Technically speaking, these texts are also plain text files (see 2.2 Plain Text Documents), and many word processing documents, image files, web sites, and other formats are primarily XML-based. For the purposes of this document, structural markup text documents include individual plain text documents written in markup languages not otherwise belonging to another format category.

### 2.3.1 SGML *with DTD/Schema*

Standard Generalized Markup Language (SGML) is a markup language used for formally describing the structure and contents of documents. It is the umbrella language under which HTML, XML, and XHTML were designed. Defined by ISO 8879:1986, SGML files use "tags" to assign style and structure to content. These tags must either be internally defined or externally defined in a document type declaration (DTD).

### 2.3.2 XML (.xml) with DTD/Schema

Extensible Markup Language (XML) is a markup language that describes a document’s storage layout and logical structure in a way that is both human and computer-readable. The term “XML” is applied to both the markup language and the documents produced with it. XML is a subset of the Standard Generalized Markup Language (SGML).

XML tags are fully extensible and user-defined. Thus, XML documents must include or refer to documentation of the meaning of the tags (markup declarations). Usually, an XML file achieves this by referencing a document type definition (DTD) or schema in its header, although the file may also include the markup declarations within the XML document itself. The Library of Congress *Sustainability of Digital Formats* database describes two types of XML documents and their markup declarations:

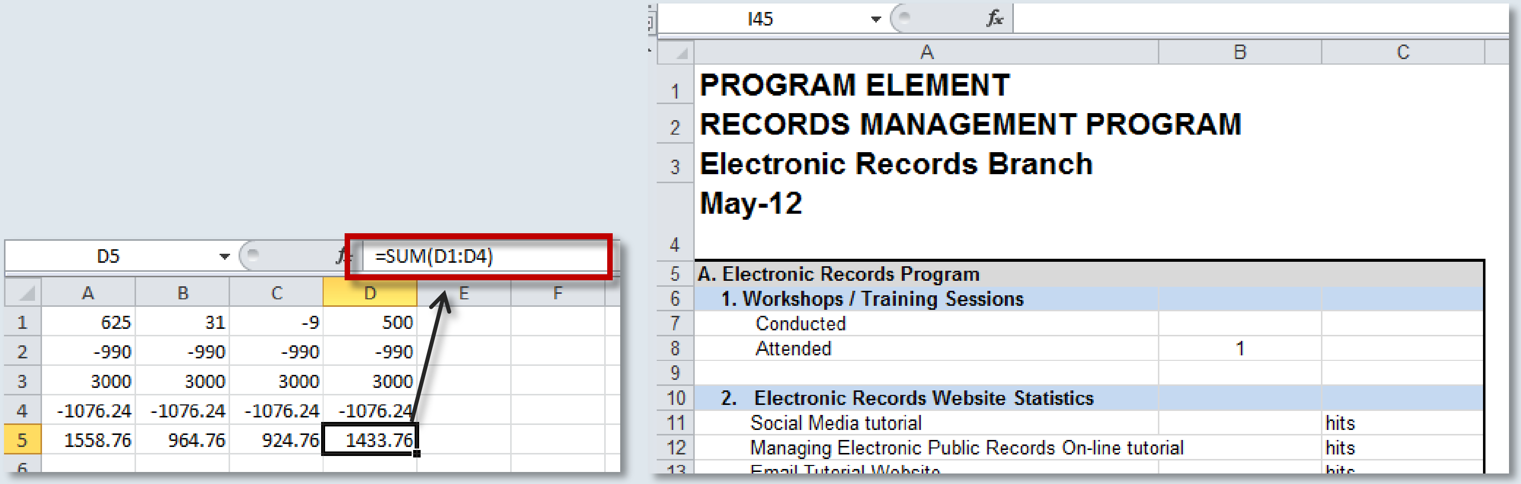
> *“XML documents fall into two broad categories: data-centric and document-centric. Data-centric documents are those where XML is used as a data transport. Examples include sales orders, patient records, directory entries, and metadata records. One significant use of data-centric XML is for manifests (lists) of digital content; another is for metadata embedded into digital content files. Document-centric documents are those in which XML is used for its SGML-like capabilities, reflecting the structure of particular classes of documents, such as books with chapters, user manuals, newsfeeds and articles incorporating explicit metadata in addition to the text. An XML document's markup structure can be defined by a schema language and validated against a definition in that language. The initial, and as of 2008, most widely used schema languages are the Document Type Definition (DTD) language and W3C XML Schema. Other schema languages exist, including RDF and RELAX-NG.”*[6]

## 2.4 Spreadsheets

Spreadsheets represent tabular data divided into columns and rows of data cells. Column and row headings identify data and allow future users to make sense and meaning of spreadsheet content. Depending on the relative importance of a spreadsheet’s content, formulas, graphs, charts, and sheets, the spreadsheet may need to be preserved in its entirety. For example, the value of cells may be created by formulae that cannot be seen if the spreadsheet is exported to PDF/A or plain text. Instead, it would need to be preserved as an OpenDocument Spreadsheet (see 2.4.1

[6] Library of Congress “XML (Extensible Markup Language,” *Sustainability of Digital Formats: Planning for Library of Congress Collections*, http://www.digitalpreservation.gov/formats/fdd/fdd000075.shtml (accessed 4/26/2012).

OpenDocument Spreadsheet ). Your agency or office will need to carefully determine whether this hidden information (or “metadata”) merits preservation.[7]



D5 =SUM(D1:D4)

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | 625 | 31 | -9 | 500 | | |
| 2 | -990 | -990 | -990 | -990 | | |
| 3 | 3000 | 3000 | 3000 | 3000 | | |
| 4 | -1076.24 | -1076.24 | -1076.24 | -1076.24 | | |
| 5 | 1558.76 | 964.76 | 924.76 | 1433.76 | | |

I45

| | A | B | C |
|---|---|---|---|
| 1 | PROGRAM ELEMENT | | |
| 2 | RECORDS MANAGEMENT PROGRAM | | |
| 3 | Electronic Records Branch | | |
| 4 | May-12 | | |
| 5 | A. Electronic Records Program | | |
| 6 | 1. Workshops / Training Sessions | | |
| 7 | Conducted | | |
| 8 | Attended | 1 | |
| 9 | | | |
| 10 | 2. Electronic Records Website Statistics | | |
| 11 | Social Media tutorial | | hits |
| 12 | Managing Electronic Public Records On-line tutorial | | hits |
| 13 | Email Tutorial Website | | hits |

**Many spreadsheets, like those pictured above, have important metadata such as formulas and styling information. This metadata is not always visible to the reader but is critical to rendering the data. When deciding between formats, it is important to consider whether your spreadsheets include this kind of information. OpenDocument Spreadsheets are capable of preserving formulas, hyperlinks, graphs, charts, and the relationships between multiple sheets. Comma-separated files and tab-delimited files are not.[6]**

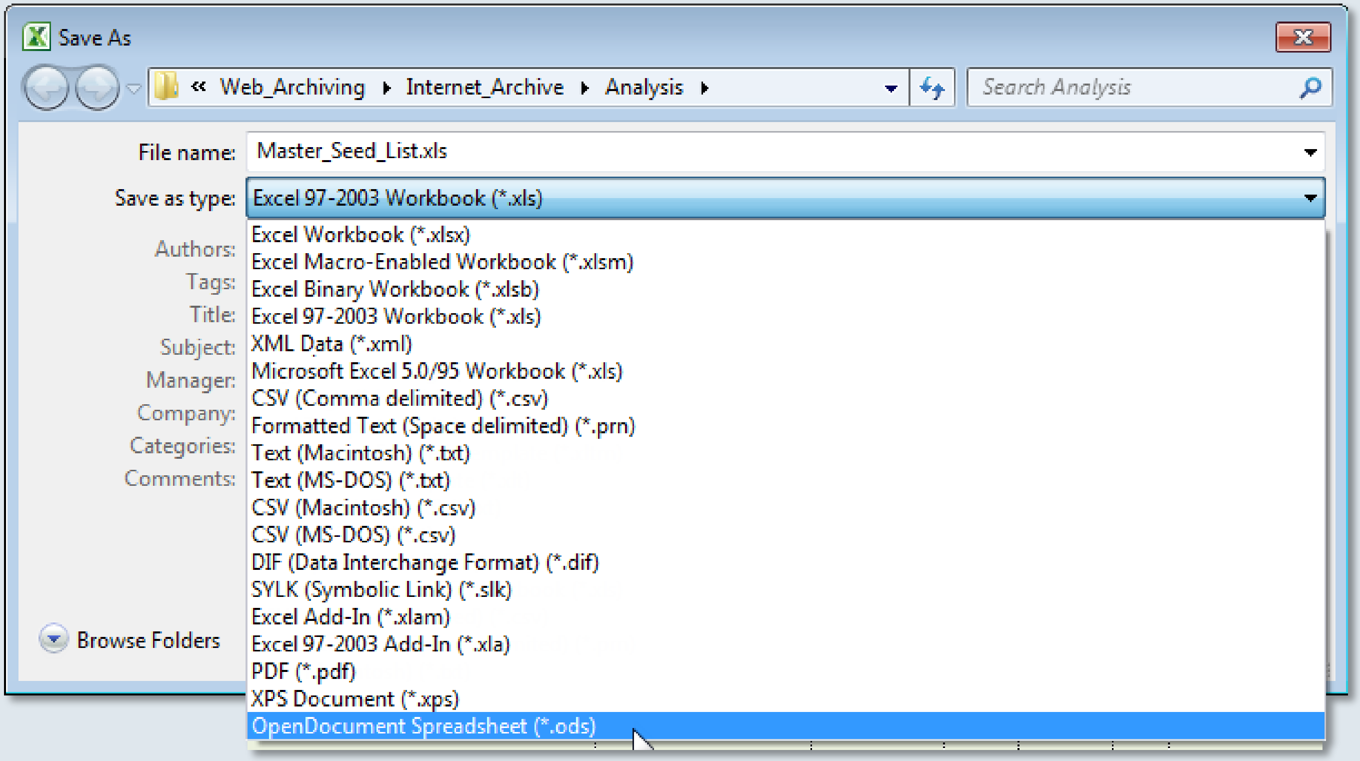
[7] For more information about retention of metadata, see *Metadata as a Public Record in North Carolina: Best Practices Guidelines for Its Retention and Disposition* (11/2010), http://www.records.ncdcr.gov/guides/Metadata_Guidelines_%2020101108.pdf (accessed 9/7/2012).

### 2.4.1 OpenDocument Spreadsheet (.ods)

OpenDocument Spreadsheet is a sub-type of the OpenDocument Format (ODF), an open source file format for spreadsheets, charts, presentations, and word processing documents. Originally created by Sun Microsystems, the current standards were developed by the Organization for the Advancement of Structured Information Standards (OASIS) Open Document Format for Office Applications committee. The format is based on the XML format used by the OpenOffice.org office suite. The format is also published (in one of its version 1.0 manifestations) as ISO/IED international standard 26300:2006. See also 2.1.2 OpenDocument Text (.odt) and 2.10.1 OpenDocument Presentation (.odp).

In almost all cases, an OpenDocument Spreadsheet "file" with the .ods extension is actually a package of several files that have been compressed into a single ZIP file package. Within the zipped package are several separate files that represent the content of the document, its styling, metadata, settings, and a manifest of the zip package files. An OpenDocument Spreadsheet file can also be a single, flat XML file; this is rare, however, and the associated file extension is usually .xml or .fods.

⚠ **The OpenDocument Spreadsheet format *does* preserve styling, formulas, graphs, charts, and the relationships between multiple sheets. If, however, you are converting your file to ODS from another format, such as a Microsoft® Excel® XLS or XLSX file, be sure to check that styling, formulas, graphs, charts, and sheet relationships were converted properly.**



File name: Master_Seed_List.xls

Save as type: Excel 97-2003 Workbook (*.xls)

Save As
« Web_Archiving ▸ Internet_Archive ▸ Analysis ▸
Search Analysis
Authors:
Tags:
Title:
Subject:
Manager:
Company:
Categories:
Comments:
Browse Folders
Excel Workbook (*.xlsx)
Excel Macro-Enabled Workbook (*.xlsm)
Excel Binary Workbook (*.xlsb)
Excel 97-2003 Workbook (*.xls)
XML Data (*.xml)
Microsoft Excel 5.0/95 Workbook (*.xls)
CSV (Comma delimited) (*.csv)
Formatted Text (Space delimited) (*.prn)
Text (Macintosh) (*.txt)
Text (MS-DOS) (*.txt)
CSV (Macintosh) (*.csv)
CSV (MS-DOS) (*.csv)
DIF (Data Interchange Format) (*.dif)
SYLK (Symbolic Link) (*.slk)
Excel Add-In (*.xlam)
Excel 97-2003 Add-In (*.xla)
PDF (*.pdf)
XPS Document (*.xps)
OpenDocument Spreadsheet (*.ods)

**Most spreadsheet editing software is capable of outputting to OpenDocument Spreadsheet format, as in this example from Microsoft® Office 2010.**

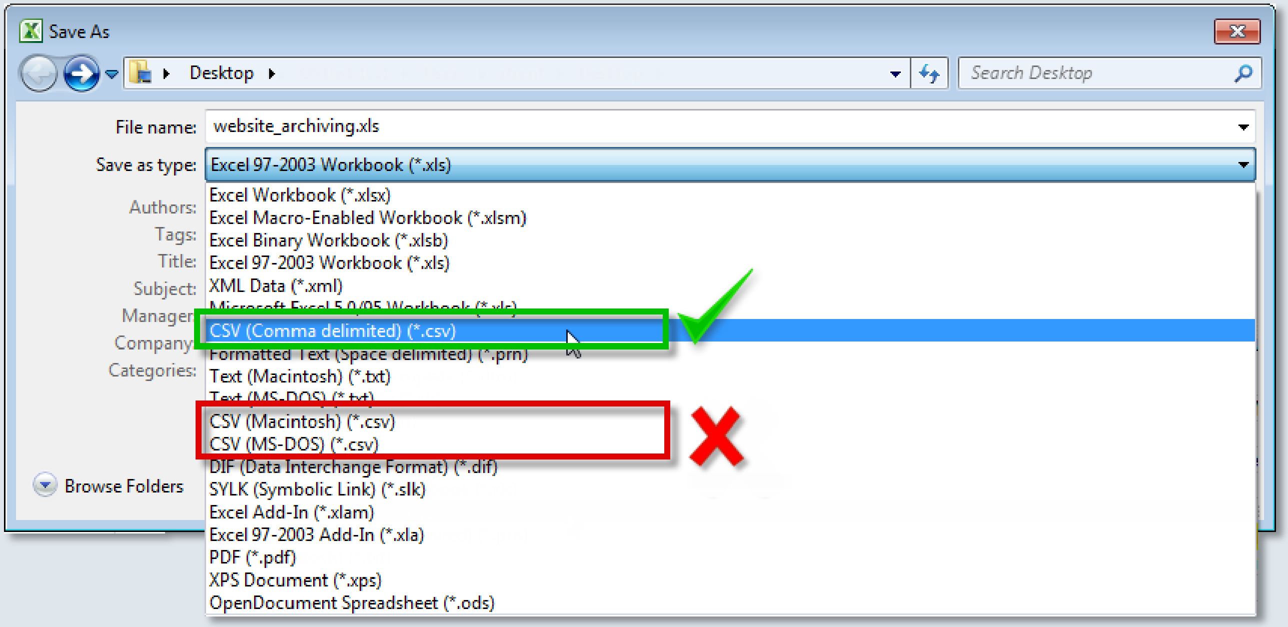
### 2.4.2 Comma-separated file (.csv)

See also 2.2.2 Comma-separated file (.csv) *US-ASCII or UTF-8 encoding*

Comma-separated files are plain text files that store tabular data. They are capable of storing spreadsheets without styling or formatting (such as borders, fonts, column widths, etc.) Like files with the .txt extension, they are usually encoded in either US-ASCII or Unicode UTF-8.

They are distinguished by the fact that they contain values separated by commas and line breaks such that spreadsheet and database applications (like Microsoft® Excel® and Access®) can easily open and parse the data.

- ⚠ **Comma-separated files cannot preserve styling, formulas, graphs, charts, or the relationships between multiple sheets. This can be important information that has been identified as having enduring value, in which case the OpenDocument Spreadsheet format should be used.**
- ⚠ **TIP: If you save from Microsoft® Excel® as a CSV file, choose "CSV (Comma delimited) (*.csv)", instead of "CSV (Macintosh) (*.csv)" or "CSV (MS-DOS) (*.csv)":**



Save As

Desktop ▸

Search Desktop

File name: website_archiving.xls

Save as type: Excel 97-2003 Workbook (*.xls)

Authors:
Tags:
Title:
Subject:
Manager
Company
Categories:

Excel Workbook (*.xlsx)
Excel Macro-Enabled Workbook (*.xlsm)
Excel Binary Workbook (*.xlsb)
Excel 97-2003 Workbook (*.xls)
XML Data (*.xml)
Microsoft Excel 5.0/95 Workbook (*.xls)
CSV (Comma delimited) (*.csv) ✓
Formatted Text (Space delimited) (*.prn)
Text (Macintosh) (*.txt)
Text (MS-DOS) (*.txt)
CSV (Macintosh) (*.csv) ✗
CSV (MS-DOS) (*.csv) ✗
DIF (Data Interchange Format) (*.dif)
SYLK (Symbolic Link) (*.slk)
Excel Add-In (*.xlam)
Excel 97-2003 Add-In (*.xla)
PDF (*.pdf)
XPS Document (*.xps)
OpenDocument Spreadsheet (*.ods)

Browse Folders

### 2.4.3 Tab-delimited file (.txt)

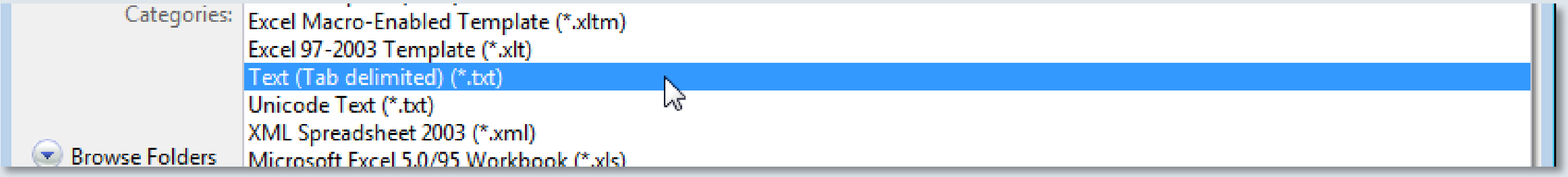
See also 2.2.3 Tab-delimited file (.txt) *US-ASCII or UTF-8 encoding*

Tab-delimited files are similar to comma separated files, the difference being that the values in one are separated by commas and in the other by tabs. Tab-delimited files carry the standard .txt extension. Like comma-separated files, tab-delimited files are not capable of storing spreadsheets formula, styling, or formatting (such as borders, fonts, column widths, etc.).

As with the comma-separated files described above, tab-delimited files should be encoded in either US-ASCII or Unicode UTF-8.

⚠ **Tab-delimited files cannot preserve styling, formulas, graphs, charts, or the relationships between multiple sheets. This can be important information that has been identified as having enduring value, in which case the OpenDocument Spreadsheet format should be used.**

⚠ **Like the ODS and CSV files described above, tab delimited files can be saved from most spreadsheet editing tools, including Microsoft® Office 2010:**


Categories:
Browse Folders
Excel Macro-Enabled Template (*.xltm)
Excel 97-2003 Template (*.xlt)
Text (Tab delimited) (*.txt)
Unicode Text (*.txt)
XML Spreadsheet 2003 (*.xml)
Microsoft Excel 5.0/95 Workbook (*.xls)

### 2.4.4 PDF/A-1a (.pdf) (ISO 19005-1 compliant PDF/A)

PDF/A may be an appropriate format for preserving spreadsheets, where styling, graphs, and charts are important elements to preserve, but formulas are not. PDF/A preserves the rendering—or "look and feel"—of the original spreadsheet, but hidden types of information like formulas are lost.

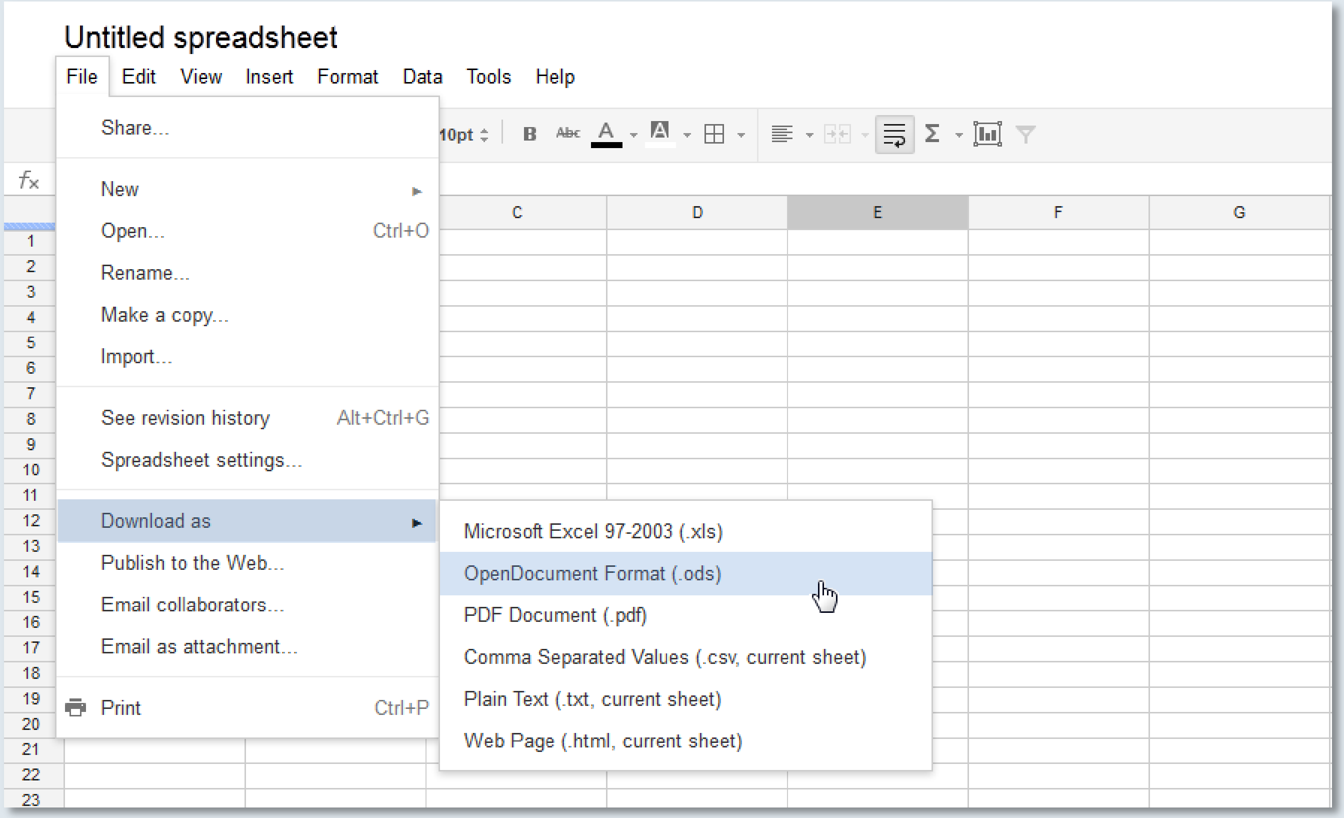
PDF-Archival (more commonly known as PDF/A) is an international standard developed by the Association for Information and Image Management International (AIIM International) for the use of PDF files for archiving and preservation of electronic documents.  The State Archives of North Carolina recommends PDF/A, Version 1, full compliance (PDF/A-1a) as a preservation format for word processing documents and other files. See also 2.1.1 PDF/A-1a (.pdf) (ISO 19005-1 compliant PDF/A).

⚠ **PDF/A files cannot preserve formulas, but can preserve styling, graphs, charts, and multiple sheets. Such elements may not render precisely the same as in the original, so it is important to check the PDF file to ensure it saved as desired.**

### *2.4.5 Special Note on Google Docs™*

Google Docs™ is a cloud-based document editing service offered by Google™. Spreadsheets may be created on Google Docs™ and exported in various formats, including Microsoft® Excel® 97-2003 (.xls), OpenDocument Spreadsheet (.ods), Comma-separated file (.csv), HTML (.html), and others. The recommendations described in this document apply to all spreadsheets, regardless of whether they were created using Google Docs™. The State Archives of North Carolina recommends that spreadsheets with styling, formulas, graphs, charts, or relationships between multiple sheets be saved in the OpenDocument Spreadsheet format. Those without styling, formulas, graphs, charts, or multiple sheets may be saved as comma-separated or tab-delimited files. See **2.4.1 OpenDocument Spreadsheets (.ods)**, **2.4.2 Comma-separated file (.csv)**, and **2.4.3 Tab-delimited file (.csv)**.

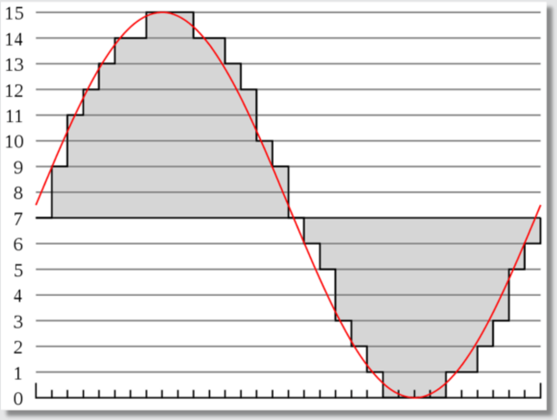
⚠ **Currently, state agencies are required to keep public records on state-owned devices and servers. Session Laws of North Carolina, SL 2011-39 §11(c) mandates that "State agencies developing and implementing information technology projects/applications shall use the State infrastructure to host their projects." Agencies may obtain an exception to this requirement. However, in the absence of such an exception, state agencies should ensure that a copy of any records created in the Google Docs™ cloud have also been exported and retained on state-owned devices. Please note that "state agencies" does not include local governments.**



**Documents created in Google Docs™ should be exported for long-term preservation as OpenDocument Format (.ods). Alternatively, documents can be exported and retained as Comma Separated Values (.csv) if there is no need to retain formulas, styling, graphs, charts, or the relationships between multiple sheets.**

## 2.5 Audio

Digitized audio “samples” sound waves at intervals, rather than recording the entire continuous sound wave as analog audio does. The digitized samples are then encoded into binary signal and packaged into a file format that tells software how to read the encoded binary data. The digitized audio file format also provides technical and descriptive information about the file (called “metadata”), such as the sampling rate, the quality of each sample (measured by bit depth), the creator of the original audio, the playback time, the date of creation, etc.



Sampling and 4-bit quantization of an analog signal (red) using Pulse-code modulation. The red line is the analog sound wave, and the gray area is the digitized approximation.

*Image courtesy of Wikipedia, [http://en.wikipedia.org/wiki/File:Pcm.svg](http://en.wikipedia.org/wiki/File:Pcm.svg) (accessed 4/27/2012)*

### 2.5.1 Broadcast WAVE Format LPCM (.wav)

The Broadcast WAVE format (BWF) with LPCM encoding is a subtype of the WAVE format (Waveform Audio File Format). In 1997, the BWF format was introduced by the European Broadcast Union (EBU) in 1997 and has since gained widespread use as the preferred archival format for audio files. Version 0 appeared in 1997, Version 1in 2001, and Version 2 in May 2011. Versions 0 and 1 are very similar, and Version 2 includes new loudness metadata.

The standard WAVE specification allows for an unlimited number of data "chunks" to sit in the head of a WAVE file. A BWF file simply includes additional metadata in the head of the file, including the EBU's "Broadcast Audio Extension" chunk, commonly known as the "bext" chunk. The bext chunk allows for important archival metadata to be embedded in the file, including the title of the recording, the recording's creator, whether the recording is part of a compilation, and much more. This information tells listeners what they are listening to, identifies essential preservation information, and allows multi-part recordings (such as multiple tracks) to be played back properly.

The data within a WAVE file is usually encoded with Linear Pulse Code Modulated Audio (LPCM), although it can also contain other variations of Pulse Code Modulated Audio (such as DPCM or ADPCM) and MPEG-encoded audio. The recommended preservation standard is to use LPCM. Alternative encodings are rarely used.

⚠ **Please contact the State Archives for more information about tools to convert simple WAVE files to BWF.**

⚠ **For more technical information about BWF preservation, please see the Federal Agencies Digitization Guidelines Initiative Audio-Visual Working Group recommendations.[8]**

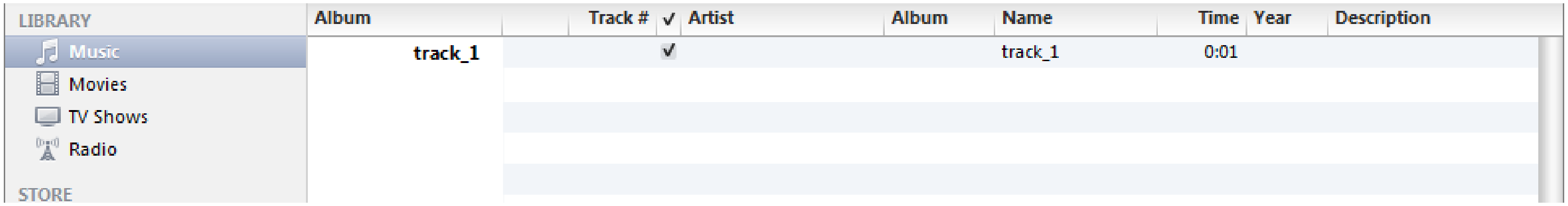
### 2.5.2 WAVE Format LPCM (.wav)

The Waveform Audio File Format (WAVE) is a standard master format for digital audio. Although it can contain compressed audio, WAVE files nearly always contain audio in uncompressed linear pulse code modulation format (LPCM).

WAVE files are widely used throughout the commercial and preservation sectors with a standardized set of additional metadata fields contained within the "bext" header chunk (see 2.5.1 Broadcast WAFE Format LPCM (.wav)). WAVE files that do not contain this additional

[8] Federal Agencies Digitization Guidelines Initiative (FADGI), *Guidelines: Embedded Metadata in Broadcast WAVE Files*, http://www.digitizationguidelines.gov/guidelines/digitize-embedding.html (accessed 9/7/2012).

metadata chunk will be missing important information that will aid in their long-term preservation, and may not easily be identifiable to listeners:



| Album | | Track # | ✓ | Artist | Album | Name | Time | Year | Description |
|---|---|---|---|---|---|---|---|---|---|
| track_1 | | | ✓ | | | track_1 | 0:01 | | |

## 2.6 Digital Video

Digital videos combine multiple elements, including visual data, audio data, subtitles or pointers to external subtitles, and descriptive information (metadata) essential for playback. Digital video files are complex, and have many layers of encoded data. In order to be able to access a digital video file, software must be able to recognize not only the umbrella file format, but also the encoders used to package the video and audio inside the file format. An MXF file, for example, may contain JPEG2000-encoded image files representing every frame in the video, wrapped into the Motion JPEG2000 format, combined with PCM audio. MXF provides the final container that links the Motion JPEG with the PCM audio, but it could also be used to link other forms of audio and video. Although the file extension (.mxf, .mov, or .mp4, for example) reflects the final container, it does not necessarily identify the component parts of the digital video.

### 2.6.1 AVI, full frame (uncompressed), WAVE PCM audio (.avi)

AVI, or Audio Video Interleaved, is a multimedia container file format developed by Microsoft®. Conforming to RIFF (Resource Interchange File Formats) AVI is a fully documented, proprietary format that has been widely adopted for video production and filmmaking. The National Archives and Records Administration (NARA) uses AVI as a preservation master format for reformatted video materials, and NARA supports the open-source AVI MetaEdit tool for the capture and normalization of AVI file embedded metadata.[9]

AVI files may contain full frame uncompressed video or compressed video, including MPEG, JPEG 2000, DV Digital Video, DivX, and other compression codecs. Audio in AVI files is WAVE PCM.

### *2.6.2 Special Note on SD (Standard Definition) and HD (High Definition) videos*

[9] AVI-MetaEdit can be downloaded from NARA's Github site: https://github.com/usnationalarchives (accessed 5/17/2012).

Several factors independent of file format help determine the quality and playability of digital video files, including the display resolution, scanning type (progressive scanning or interlaced scanning), and frame rate. The State Archives will accept digital video files that adhere to established NTSC standard broadcast resolutions for either SD (Standard Definition) or HD (High Definition) video:[10]

**Standard Definition NTSC:**
720 x 480 29.97fps (480i, 480p)
Aspect ratios: 4:3 or 16:9

**High Definition NTSC:**
1280 x 720 (720p60, 720p30, 720p24)
1920 x 1080 (1080i60, 1080p30, 1080p24)
Aspect ratio: 16:9

## 2.7 Raster Images

Raster images, also known as "bit-mapped" images or "bitmaps," are still images created with a grid of pixels, or very small squares of color. The following is a useful introduction to raster images from the Library and Archives Canada (LAC):[11]

> *A raster image is comprised of bits of information representing uniquely valued pixels in the form of a grid. Image resolution is measured by pixels per inch (PPI); however the printing abbreviation DPI (dots per inch) is also commonly used to describe image resolution. All digital photographs, regardless of file type, are raster images.*
>
> *The more pixels there are in relation to the area, the higher the resolution. The higher the resolution, the sharper the image is and the larger the file. […]*
>
> *Digital image resolution is greatly misunderstood. Digital images themselves have no size other than the number of pixels they contain. The image only has real dimensions (inches or cm) when it is in an analogue form before digitization, or after it has been printed.*

[10] See similar recommendations from Library and Archives Canada (LAC), *Local Digital Format Registry (LDFR): File Format Guidelines for Preservation and Long-term Storage, Version 1.0* (2010), 21-22.

[11] Library and Archives Canada (LAC), *Local Digital Format Registry (LDFR): File Format Guidelines for Preservation and Long-term Storage, Version 1.0* (2010), 24-25.

*There are two basic measures for digital imagery characteristics:*

- *Spatial resolution – capturing detail (PPI) and,*
- *Tonal resolution – colour, bit depth and dynamic range.*

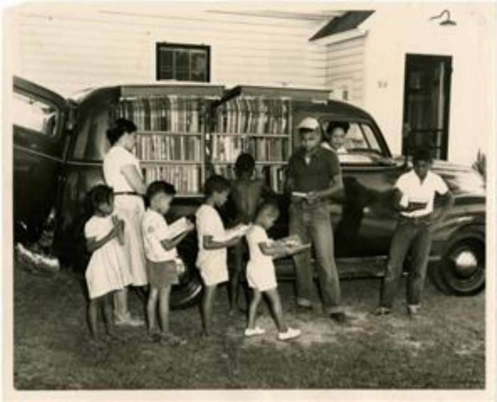
*Generally, the higher the PPI and the larger the bit depth, the more accurate the image will be to its original colour. Black and white images are not characterized by colour resolution. They are comprised of brightness values that represent 256 different shades of gray.*

| *Colour Depth* | *Number of Colours Visible* |
|---|---|
| *1 bit (monochrome)* | *2* |
| *4 bit* | *16* |
| *8 bit (indexed colour)* | *256* |
| *24 bit (true colour)* | *16,777,216* |

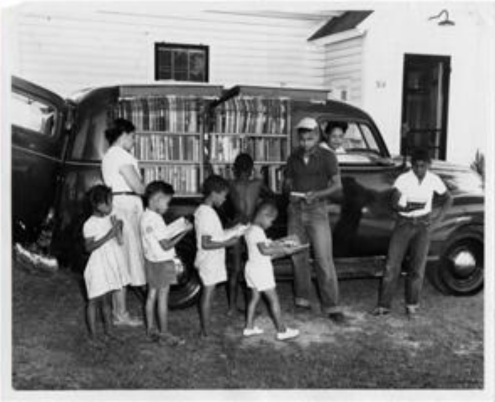
*Common "colour resolutions" are 1 bit per pixel, for solid black-and-white nonrealistic images; 8 bits per pixel for grayscale images, nonrealistic colour images, and coarse realistic images; and 24 bits per pixel, for "photographic quality" realistic images. 48 bits per pixel is in increasing use for ultrahigh quality images.*

*Grayscale images have a maximum colour depth of 8 bits. This is because when defining shades of gray in terms of RGB, each of the 3 red, green and blue components must be equal (i.e. R=192 G=192 B=192, or R=128 G=128 B=128). Since these three components must be equal, there are only 256 possible combinations, which equals 8 bits of colour.*

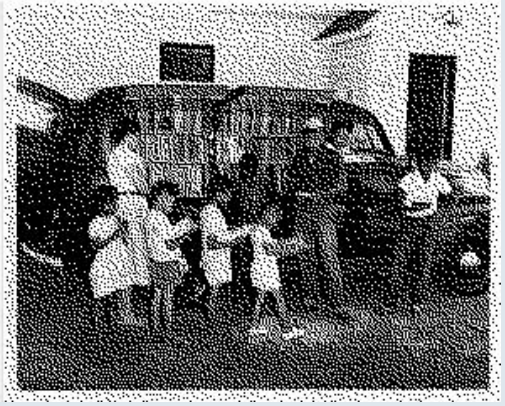
*Indexed colour images are limited to a maximum of 256 colours (8-bit), which can be any 256 colours from the set of 16.7 million 24 bit colours. Each image file contains [its] own palette which provides a reference index number used by the computer to identify each colour.*



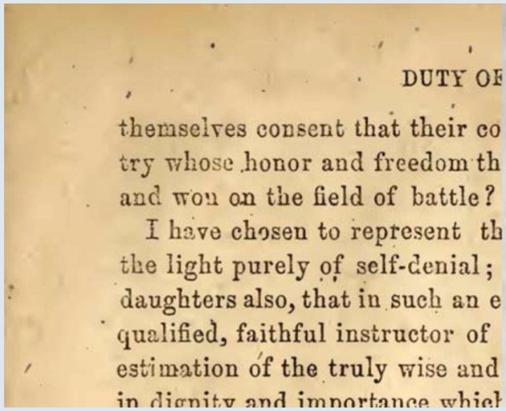
RGB full-color photographic image



Grayscale photographic image



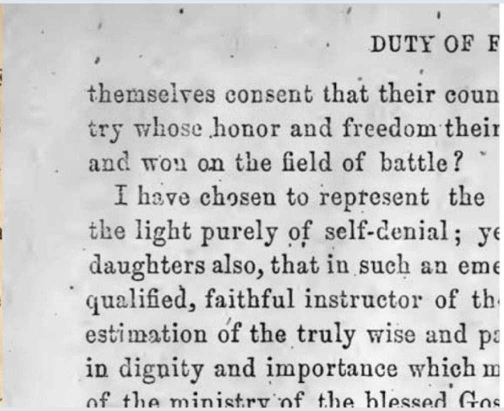
Bi-tonal photographic image[12]



DUTY OF

themselves consent that their co
try whose honor and freedom th
and won on the field of battle?
I have chosen to represent th
the light purely of self-denial;
daughters also, that in such an e
qualified, faithful instructor of
estimation of the truly wise and
in dignity and importance whicl

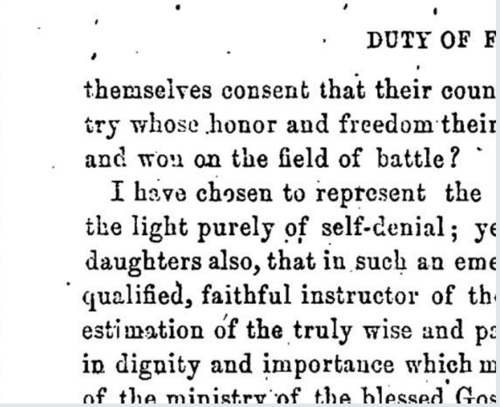
RGB full-color textual image



DUTY OF F

themselves consent that their coun
try whose honor and freedom their
and won on the field of battle?
I have chosen to represent the
the light purely of self-denial; ye
daughters also, that in such an eme
qualified, faithful instructor of th
estimation of the truly wise and pa
in dignity and importance which m
of the ministry of the blessed Gos

Grayscale textual image



DUTY OF F

themselves consent that their coun
try whose honor and freedom their
and won on the field of battle?
I have chosen to represent the
the light purely of self-denial; ye
daughters also, that in such an eme
qualified, faithful instructor of th
estimation of the truly wise and pa
in dignity and importance which m
of the ministry of the blessed Gos

Bi-tonal textual image[13]

[12] [African-American children line up outside of Albemarle Region bookmobile], Photograph, Public Library History Files, State Library of North Carolina, North Carolina Digital Collections, http://digital.ncdcr.gov/cdm4/item_viewer.php?CISOROOT=/p249901coll36&CISOPTR=195&CISOBOX=1&REC=5 (accessed 4/30/2012).

### 2.7.1 TIFF (.tif), uncompressed

The Tagged Image File Format (TIFF) is the preferred preservation file format for raster images. Although the TIFF specification is owned by Adobe® Corporation, the format is fully documented, extensible, and widely adopted.[14] The State Archives recommends the use of **TIFF v. 6.0 uncompressed baseline RGB** for color images, **TIFF v. 6.0 uncompressed baseline grayscale** for grayscale images, and **TIFF v. 6.0 Group IV/Huffman compressed baseline bi-tonal** for typographic documents where there are no fine details, light markings, handwritten or pencil notations. This means that preservation TIFF files should be:

1. **version 6.0**, which was released in 1992 and is the most recent TIFF specification. TIFF 6.0 Supplement 2, which was released in 2002 and introduced two additional compression schemes used when saving TIFF files in Adobe® Photoshop®, does not affect the recommended preservation format.[15]
2. **baseline.** This simply means that the minimum (or baseline) tags are present that make a TIFF file a TIFF file. If your file does not meet the minimum baseline tagging requirements, it is not a valid TIFF file and software may report that the file is corrupted when attempting to open it. Software will likely be able to save TIFF files at baseline by default.
3. **RGB, grayscale, or bi-tonal**, depending on the appearance of the analog original or encoding of the born-digital original. Baseline TIFF files have four configurations: bi-tonal (black and white), grayscale, palette-color (limited color palette), and full-color RGB. If your original records are color images (photos, maps, text with colored notations, or any born-digital image that originated in RGB), you should preserve that image as a full-color RGB TIFF file. If you are scanning grayscale documents (such as grayscale maps, typographic documents that are difficult to read or that have pencil markings, handwritten documents, etc.) or have born-digital images that originated in grayscale, you should save the images as uncompressed grayscale TIFF files. For scanned images of black and white typographic documents where visual detail is not important and there are no fine details, light markings, handwritten or pencil notations,  Group IV/Huffman bi-tonal (black and white) TIFF files may be used.

[13] "The Duty of Females in Relation to the Future Educational Interests of Our Country," *North Carolina Journal of Education,* vol. 7 no. 4 (July 1864): 91, http://digital.ncdcr.gov/cdm4/document.php?CISOROOT=/p249901coll37&CISOPTR=14329&REC=1 (accessed 4/30/2012).
[14] *TIFF Revision 6.0*, June 3, 1992 http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf (access 4/30/2012).
[15] *Adobe Photoshop® TIFF Technical Notes*, March 22, 2002 http://partners.adobe.com/public/developer/en/tiff/TIFFphotoshop.pdf (accessed 4/30/2012).

| Recommended TIFF file format | Examples of original formats |
|---|---|
| TIFF v. 6.0 **uncompressed** baseline **RGB** | all born-digital color images,<br>digitized color photographs,<br>grayscale photographs where discoloration or sepia tone are of historic value,<br>text with colored ink, paper, highlighting or annotations of value,<br>color maps |
| TIFF v. 6.0 **uncompressed** baseline **grayscale** | all born-digital grayscale images,<br>grayscale photographs without discoloration or sepia tone of historic value,<br>handwritten text with no color of value,<br>typographic text that is difficult to read in bi-tonal (black and white) scans |
| TIFF v. 6.0 **Group IV/Huffman compressed** baseline **bi-tonal** | typographic text with clear printing where text is not fine or faint |

### 2.7.2 JPEG 2000 (.jp2)

Joint Photographic Experts Group JPEG 2000 is an open, published ISO standard (ISO 15444-1:2004). TIFF has long been established as the archival preservation file format of choice for raster images, and JPEG 2000 is increasingly being considered a viable format as well. It is yet to be widely adopted, however, and **the State Archives recommends JPEG 2000 with reservations. Agencies should use JPEG 2000 as a preservation format only where staff with technical proficiency are familiar with the format and/or where JPEG 2000 is already in use in the office.** Although JPEG 2000 is less widely adopted than TIFF, JPEG 2000 offers several advantages, including a highly efficient lossless encoding that allows for very high quality images at very low file sizes. There are three types of JPEG 2000 image file formats, the first of which is the most widely accepted for long-term preservation:

- **JPEG 2000 Part 1 (JP2, .jp2)** — International ISO standard ISO 15444-1:2004, JP2 is the core coding system and the most widely adopted by preservation institutions, including Library of Congress (LOC), Library and Archives Canada (LAC), the National Library of the Netherlands, the British Library, the Wellcome Library, the National Library of Norway, and the National Library of the Czech Republic.[16]

- **JPEG 2000 Part 2 (JPX, .jpx, .jpf) —** International ISO standard ISO 15444-2:2001, JPX is an extension of JP2 that allows for additional colorspaces, the specification of opacity, standardized metadata, multiple image data streams, and more non-contiguous internal organization of the image data.[17]Although the official MIME type extension is .jpf, some applications may save files as .jpx.[18] Library of Congress notes the following information in its recommendation of JPX alongside JP2 as a preservation format for still images: "The JPX level of the JPEG2000 standard supports more effective color management than the level 1 format (JP2). JPEG2000 offers many options for choices of quality level, and storage order for the encoded image data (codestream). Future investigation is needed to determine whether particular options should be encouraged or avoided when the objective is responsible long-term custody."[19]

[16] "JPEG 2000 Part 1 (Core) jp2 File Format," *Sustainability of Digital Formats: Planning for Library of Congress Collections*, http://www.digitalpreservation.gov/formats/fdd/fdd000143.shtml (accessed 05/02/2012); [16] Library and Archives Canada (LAC), *Local Digital Format Registry (LDFR): File Format Guidelines for Preservation and Long-term Storage, Version 1.0* (2010); R. Gillesse, J. Rog, & A. Verheusen, *Alternative File Formats for Storing Master Images of Digitisation Projects* (Den Haag: Koninklijke Bibliotheek, 2008) http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/Alternative_File_Formats_for_Storing_Masters_2_1.pdf (accessed 5/2/2012); R. McLeod & P. Wheatley, *Preservation Plan for Microsoft — Update Digital Preservation Team* (London: British Library, 2007) http://www.bl.uk/aboutus/stratpolprog/ccare/introduction/digital/digpresmicro.pdf (accessed 5/2/2012); C. Henshaw, *ICC profiles and LuraWave* (London: Wellcome Library, 2010) jpeg2000wellcomelibrary.blogspot.com/2011/05/icc-profiles-and-lurawave.html (accessed 5/2/2012); National Library of Norway, *Digitization of books in the National Library — methodology and lessons learned* (Oslo: National Library of Norway, 2007) http://www.nb.no/content/download/2326/18198/version/1/file/digitizing-books_sep07.pdf (accessed 5/2/2012); B. Vychodil, "JPEG2000 - Specifications for The National Library of the Czech Republic," *Seminar JPEG 2000 for the Practitioner* (London: Wellcome Trust, 16 Nov 2010) http://www.dpconline.org/component/docman/doc_download/520-jp2knov2010bedrich (accessed 5/2/2012).

[17] "JPEG2000-hul Module," *Jhove* http://hul.harvard.edu/jhove/jpeg2000-hul.html (accessed 5/2/2012).

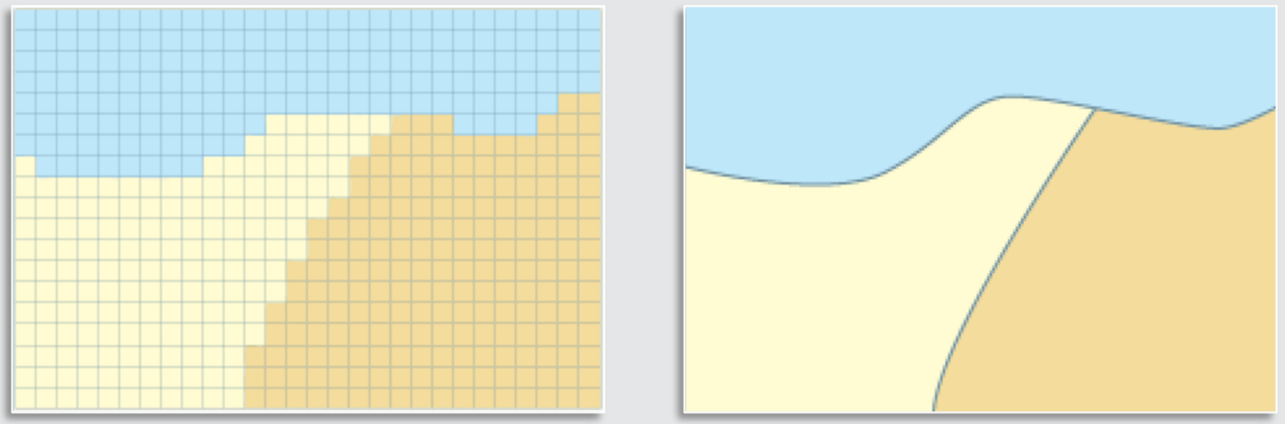
[18] D. Singer, et al., "MIME Type Registrations for JPEG 2000 (ISO/IEC 15444)," April 2004 http://www.ietf.org/rfc/rfc3745.txt (accessed 5/2/2012).

[19] Table 2, Note 2, "Still Image: Curator's View," *Sustainability of Digital Formats: Planning for Library of Congress Collections*, http://digitalpreservation.gov/formats/content/still_curator.shtml (accessed 05/02/2012).

- **JPEG 2000 Part 6 (JPM, .jpm) —** International ISO standard ICO 15444-6:2003, and based on the Mixed Raster Content standard ICO/IEC 16485:2000, JPM is designed to combine bit-tonal and continuous-tone images into compound images. Library and Archives Canada (LAC) includes this format in their description of the recommended JPEG 2000 format.[20]

## 2.8 Vector Images

Whereas raster images are built by small dots of color, vector images are created mathematically with the geometry of points, lines, curves, and polygons. Raster images tend to be used for photographs and photo-realistic images, while vector images are used for structured pictures, such as architectural drawings, graphic designs, and engineering drawings. Vector images are also used widely in geospatial databases, which this section does not cover. For geospatial vector sets, see 2.13 Geospatial Vector Datasets.



RASTER REPRESENTATION

VECTOR REPRESENTATION

*Image courtesy of Library and Archives Canada (LAC)*, Local Digital Format Registry (LDFR): File Format Guidelines for Preservation and Long-term Storage, Version 1.0 *(2010), 25.*

[20] Library and Archives Canada (LAC), *Local Digital Format Registry (LDFR): File Format Guidelines for Preservation and Long-term Storage, Version 1.0* (2010), 26.

### 2.8.1 Scalable Vector Graphics 1.1 (.svg)

SVG is a widely adopted and open standard from W3C that is used for creating two-dimensional graphics in XML.[21] Its official description is as follows: “SVG is a language for describing two-dimensional graphics in XML [XML10]. SVG allows for three types of graphic objects: vector graphic shapes (e.g., paths consisting of straight lines and curves), images and text. Graphical objects can be grouped, styled, transformed and composited into previously rendered objects. The feature set includes nested transformations, clipping paths, alpha masks, filter effects and template objects.”[22] The National Archives of the UK writes,

> *“SVG supports 24-bit colour, and allows the creation of sophisticated dynamic and interactive graphics. Being entirely XML-based, it enjoys the advantages of extensibility, interoperability and flexibility. SVG can be easily manipulated and transformed using standard XML tools.*
>
> *SVG is an open, non-proprietary format. It is rapidly becoming a major standard for vector imagery, particularly on the Internet. It is widely supported by all the major vendors, and a number of free viewers are available for off/online viewing.”*[23]

### 2.8.2 AutoCAD® Drawing Interchange Format (.dxf)

The Drawing Interchange Format was developed and is owned by Autodesk®, the producer of AutoCAD®. AutoCAD®’s native Drawing Format, DWG, is currently the *de facto* standard for vector graphics. The DWG format, however, is proprietary and has not been released. Autodesk® instead recommends the use of the Digital Interchange Format DXF for data exchanges. The DXF specification, which is revised alongside each release of AutoCAD®, was designed to be exchanged with other CAD applications. The format is owned by Autodesk® but freely available to use. The UK National Archives writes that “DXF is a complex format, and the quality and sophistication of its implementation in different applications varies considerably. The frequent changes to the specification can also cause compatibility problems. In particular, users must be aware that some applications may read a DXF file whilst skipping unsupported features. This can lead to the loss of information in a manner that may not be obvious to the user.”[24] **The State Archives of North Carolina recommends that the AutoCAD® Drawing Interchange Format (.dxf) be used as a preservation format only where PDF/A or SVG are not appropriate.**

[21] “Scalable Vector Graphics (SVG),” W3C http://www.w3.org/Graphics/SVG/ (accessed 5/2/2012).

[22] “Scalable Vector Graphics (SVG) 1.1 (Second Edition),” W3C (August 13, 2011) http://www.w3.org/TR/SVG11/ (accessed 5/2/2012).

[23] The National Archives, *Graphics File Formats: Digital Preservation Note 4* (August, 2008), 13 http://www.nationalarchives.gov.uk/documents/graphic-file-formats.pdf (accessed 5/2/2012).

[24] The National Archives, *Graphics File Formats: Digital Preservation Note 4* (August, 2008), 11-12 http://www.nationalarchives.gov.uk/documents/graphic-file-formats.pdf (accessed 5/2/2012).

### 2.8.3 PDF/A-1a (.pdf) (ISO 19005-1 compliant PDF/A)

PDF-Archival (more commonly known as PDF/A) is an international standard developed by the Association for Information and Image Management International (AIIM International) for the use of PDF files for archiving and preservation of electronic documents. The State Archives of North Carolina recommends PDF/A, Version 1, full compliance (PDF/A-1a) as a preservation format for word processing documents and other files. For some images created in raster form, PDF/A-1a may be an appropriate preservation format, particularly for simple images where 2D visual rendering is more important than manipulability.[25] See also 2.1.1 PDF/A-1a (.pdf) (ISO 19005-1 compliant PDF/A).

## 2.9 Databases

Databases contain structured data organized into fixed fields in computer-readable files. Today, nearly all databases are relational databases, in which data is contained in a series of formally-described and related tables from which data can easily be queried. A database management system (DBMS) refers to the software used to manage databases, rather than to the content of the databases themselves. Common DBMSs include Microsoft® Access®, Microsoft® SQL Server, and Oracle® MySQL. Specialized fields like astronomy, social science, and engineering often utilize unique DBMSs. The database preservation category focuses not on the DBMSs, but on the formats designed for the exchange of data from one DBMS to another.

### 2.9.1 Software Independent Archiving of Relational Databases (SIARD)

The Software Independent Archiving of Relational Databases (SIARD) format was developed by the Swiss Federal Archives in response to the lack of a standardized archiving format for databases.[26] It is an XML-based format designed for the long-term preservation of relational database content. First introduced by the Swiss Federal Archives in 2004, it has since been further developed within the PLANETS project. In 2008, the Swiss Federal Archives released a full-fledged version of the SIARD format with associated software.[27]

[25] For more information about PDF/A as a preservation format for vector images, see David Duce, Bob Hopgood, Mike Coyne, Mike Stapleton, and George Mallen, "SVG and the Preservation of Vector Images," Digital Preservation and Records Management Programme of JISC, the Joint Information Systems Committee of the UK Higher Education ([2008]) http://www.svgopen.org/2008/papers/40-SVG_and_the_Preservation_of_Vector_Images/#d4e396 (accessed 5/8/2012); Mike Coyne, David Duce, Bob Hopgood, George Mallen, Mike Stapleton, *Study on the Significant Properties of Vector Images,* JISC Digital Preservation Programme (11/27/2007) http://www.jisc.ac.uk/media/documents/programmes/preservation/vector_images.pdf (accessed 5/8/2012).

[26] Swiss Federal Archives, "SIARD Format Description" (09/30/2008) http://www.bar.admin.ch/dienstleistungen/00823/00825/index.html?lang=en&download=NHzLpZeg7t,lnp6I0NTU042l2Z6ln1ad1IZn4Z2qZpnO2Yuq2Z6gpJCDdIR8fmym162epYbg2c_JjKbNoKSn6A and http://www.bar.admin.ch/dienstleistungen/00823/00825/index.html?lang=en (accessed 5/8/2012).

[27] Amir Bernstein, "Data Preservation: The International Challenge and the Swiss Solution," undated briefing paper, Swiss Federal Archives http://www.digitalpreservationeurope.eu/publications/briefs/database_preservation.pdf (accessed 5/8/2012).

A SIARD archive is a single uncompressed ZIP container that holds two folders: a metadata folder and a content folder. The metadata folder contains an identification of the database, format version, lists of tables, views, routines, table constraints and triggers, SQL type, LOBs (Large Objects) names, and relations. The content folder holds the schema and table data in XML files.[28]

### 2.9.2 Delimited Flat File (Plain Text) with DDL

Delimited plain text files may be used for archiving simple database content or content from legacy database applications (see **2.2.1 Plain Text** for acceptable encodings). Data fields should be delimited using commas, tabs, or another delimiter (see **2.2.1 Comma-separated file (.csv)** and **2.2.2 Tab-delimited file (.csv)**), rather than being stored as fixed-length flat files.

In order that the data be identified and made comprehensible, at minimum there must be a data definition language (DDL) accompanying the database. Any additional contextual information transferred to the State Archiving accompanying the database (such as data dictionaries and relational diagrams) should be submitted in an appropriate preservation format.

## 2.10 Presentations

Presentations are image, text, and audio-based displays of information, usually in the form of a slide show. Common tools like Microsoft® PowerPoint®, OpenOffice.org Impress, Corel® Presentations, and Google Docs™ allow users to create, edit, and present such files. Presentations may include not only images, text, and audio, but also timed animations, hyperlinks, and click effects. Web-based applications like Google Docs™ allow for the online, collaborative creation of such presentations, and export files in common pre-existing formats.

Although zooming presentations (or ZUIs, for zoom user interface) are available through tools like Prezi, these are not considered here within the scope of presentation file formats.

### 2.10.1 OpenDocument Presentation (.odp)

OpenDocument Presentation (.odp) is a sub-type of the OpenDocument Format (ODF), an open source file format for spreadsheets, charts, presentations, and word processing documents. Originally created by Sun Microsystems, the current standards were developed by the Organization for the Advancement of Structured Information Standards (OASIS) Open Document Format for Office Applications committee. The format is based on the XML format used by the OpenOffice.org office suite. The format is also published (in one of its version 1.0 manifestations) as ISO/IED international standard 26300:2006. See also **2.1.2 OpenDocument Text (.odt)** and **2.4.1 OpenDocument Spreadsheet (.ods)**.

[28] Ibid.

An OpenDocument Presentation “file” with the .odp extension is actually a package of several files that have been compressed into a single ZIP file package. Within the zipped package are several separate files that represent the content of the document (including images, notes, and text), animations and click effects, themes, styles, layouts—as well as a manifest of the zip package files.

### 2.10.2 PDF/A-1a (.pdf) (ISO 19005-1 compliant PDF/A) *for presentations without animation*

PDF-Archival (more commonly known as PDF/A) is an international standard developed by the Association for Information and Image Management International (AIIM International) for the use of PDF files for archiving and preservation of electronic documents. The State Archives of North Carolina recommends PDF/A, Version 1, full compliance (PDF/A-1a) as a preservation format for presentations without audio or animations. See 2.1.1 PDF/A-1a (.pdf) (ISO 19005-1 compliant PDF/A).

## 2.11 Email

### State Agency Employees

Pursuant to E.O. 18, Executive Branch state employees are required to retain all emails for ten years. Most agencies under E.O. 18 currently use a service called Mimosa, which automatically retains all email messages for ten years. Mimosa, however, does not fulfill state agencies’ records retention requirements. Although Mimosa automatically retains emails for a period of ten years, state agencies are also required to consult their records retention schedules and identify any emails with retention periods longer than ten years. When consulting records retention schedules, state agency employees should be aware that records are defined by content, not media. Thus, excepting the general state agency schedule, schedules will not include a record series titled simply “email.” Rather, email can contain several types of records. Examples include correspondence, meeting agendas, employee leave requests, conference materials, employee vehicle requests, law enforcement case files, and more.

Currently, emails of state employees are delivered to the State Archives of North Carolina in the Microsoft® Outlook® Personal Storage Format (.pst), where they are then converted into XML for long-term preservation. The file format recommendations below are intended for local government employees.

### Local Government Employees

Unlike state agency employees, local government employees are not affected by Executive Order 18, also known as E.O. 18. This executive order requires state agencies to retain email for ten years. Local government employees should consult their retention schedules to determine the retention periods for their emails. When consulting records retention schedules, local government employees should be aware that records are defined by content, not media. Thus, no schedule will not contain a record series titled simply “email.” Rather, email can contain several types of records.

Examples include correspondence, meeting agendas, employee leave requests, conference materials, employee vehicle requests, law enforcement case files, and more. The following format recommendations are intended for local government employees.

⚠ **NOTE: Local government employees are not subject to E.O. 18, which requires that most state agency representatives retain email for ten years.**

⚠ **NOTE: Most county and municipal email messages may be "destroyed in office after administrative value ends." "Administrative value ends" generally means "when you don't need them to do your job anymore."**

### Formats: Multiple Emails & Email Accounts

In common usage, "email" is used to refer to one of two things: (1) an individual email message or (2) an entire email account, including messages, message folders, contacts, calendars, and tasks. Most email file formats are designed to either hold a single message, multiple messages, or an entire email account, including non-message data. This section discusses file formats that aggregate multiple email messages or contain entire accounts with non-message data.

Agencies may consider retaining email in either form, depending on what records are contained within the email. Public records with long-term retention may appear in several areas throughout an email account. For example, the Governor's daily and monthly schedule file (permanent retention) might include an email calendar; county legal correspondence files (retain for 3 years after resolution) might include meeting invitations in an email calendar; municipal public relations file (retain 5 years) may include email messages arranged into folders such that the folders are valuable to retain as well.

⚠ **NOTE: Not all email clients allow entire accounts or aggregated messages to be exported. Certain email accounts may require that messages be individually.**

### 2.11.1 Microsoft® Outlook® Personal Storage Table (.pst)

The Personal Storage Table (.pst) format is an open, proprietary format owned by Microsoft® and used primarily by the Microsoft® Exchange Client, Windows® Messaging, and Microsoft® Outlook®. Although Microsoft® owns copyright to the format, it is freely published to allow open development of tools that can open, process, manage, and convert .pst files.[29]

The State Archives of North Carolina recommends that local agencies using Outlook® and needing to retain entire email accounts should retain those accounts as PST files. PST files, however, are frequently updated and should regularly be updated with new releases of Microsoft® Office. If individual employees are retaining PST files locally, each time Microsoft® Office is updated on employees' computers, the employees should open all PST files in Outlook® and re-save in the updated PST format. Keep in mind, there may be little apparent difference between two versions of a PST file. It is essential that PST files be updated in this manner in order to be retained long-term.

PST files are highly complex. They contain messages within folder hierarchies, message attachments, as well as calendars, contacts, tasks, email flags and categorization, and other data. Email accounts can quickly become very large, and PST files (like any file) have a maximum size beyond which the file may easily corrupt. PST files produced prior to Outlook® 2003 have a maximum file size of 2 GB. Those produced in Outlook® 2003 and 2007 have a default maximum size of 20 GB, and those in Outlook® 2012 50 GB. Files beyond these size limits should be divided into smaller files that do not exceed the maximum recommended size. This can be accomplished by opening the PST file in Outlook® and re-saving selected sections separately.

### 2.11.2 MBOX (.mbox, .mbx)

For those agencies that do not use Microsoft® Outlook®, MBOX may be an acceptable format for retention. The State Archives only recommends MBOX be used where PST files are not available and it is not viable to save messages individually.

MBOX is not a file format per se, but rather a family of four related storage formats. Different email clients implement MBOX in different ways, but MBOX files generally store all messages within a single email folder Different MBOX formats mark the end of one message and the beginning of another in slightly different ways. MBOX has become a de facto standard across email clients, with different clients employing one of the four types of MBOX file formats. It may be necessary to open your MBOX files in Notepad or another text editor to determine the precise MBOX format employed by your email client.

**Programs like Aid4Mail, Emailchemy or Xena can help distinguish and convert MBOX formats.**

[29] *Outlook Personal Folders (.pst) File Format Specification* http://msdn.microsoft.com/en-us/library/ff385210.aspx (accessed 5/9/2012).

It is also important to note that MBOX files include attachments (spreadsheets, word processing documents, images, etc. that the ender attached to the email) embedded into the email.[30] Note, however, that the State Archives of North Carolina recommends that any documents attached to emails be preserved separately and in their native file format. Attachments should be downloaded and preserved as separate files.

| MBOX format | File Extention | Notes |
|---|---|---|
| mboxo | .mbox<br>.mbx | Used by Eudora<br>⚠ NOTE: mboxo suffers from significant flaws that easily lead files to be unreadable. Do not use the mboxo format for email preservation. Programs like Aid4Mail, Emailchemy or Xena can help distinguish MBOX formats. |
| mboxrd | .mbox<br>.mbx | Used by Mozilla, Thunderbird, qmail |
| mboxcl | .mbox<br>.mbx | mboxcl format is like mboxo format, but includes a Content-Length field with the number of bytes in the message.[31] |
| mbocxl2 | .mbox<br>.mbx | mboxcl2 format is like mboxcl but has no >From quoting. These formats are used by SVR4 mailers. mboxcl2 cannot be read safely by mboxrd readers. |

## Formats: Individual Email Messages

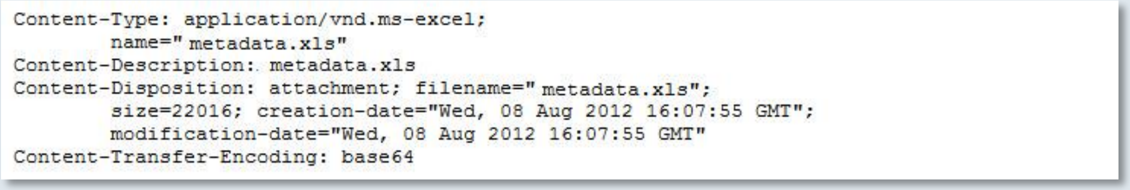
In some cases, clients may not allow users to export accounts or multiple messages into a single file. Instead, the email client may allow individual messages to be exported one-by-one. In this case, users should be careful to select a format that (1) includes as much metadata as possible and (2) which is least likely to become inaccessible in the near future.

[30] In MIME format.

[31] Mbox format definition, http://qmail.org/man/man5/mbox.html (accessed 8/2/2012).

Many email clients will offer users multiple format options for export. These email formats are not uniformly identified across email clients, and many formats are interpreted differently by different clients. Thus, the State Archives of North Carolina makes the following general guidelines for local governments, rather than recommending specific file formats:

1. Attachments should be saved as separate files, in their original format. The email message should indicate whether there are any attachments and include the filenames of those attachments. Depending on the file format, the attachment may also be embedded in the message file itself. Regardless of whether the attachment is embedded in the message, it should also be saved separately in its original format to ensure that it can be opened at a later date.



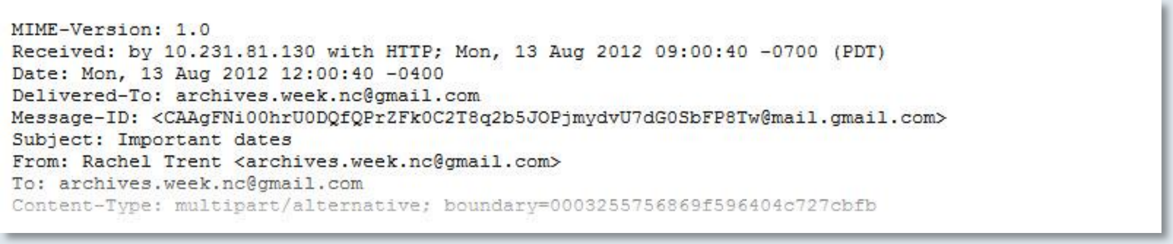
```
Content-Type: application/vnd.ms-excel;
        name=" metadata.xls"
Content-Description: metadata.xls
Content-Disposition: attachment; filename=" metadata.xls";
        size=22016; creation-date="Wed, 08 Aug 2012 16:07:55 GMT";
        modification-date="Wed, 08 Aug 2012 16:07:55 GMT"
Content-Transfer-Encoding: base64
```

**Shown here is a portion of an email (saved in plain text) that has a Microsoft® Excel® file attached. Note that in this example, the filename, size, and timestamps are preserved.**

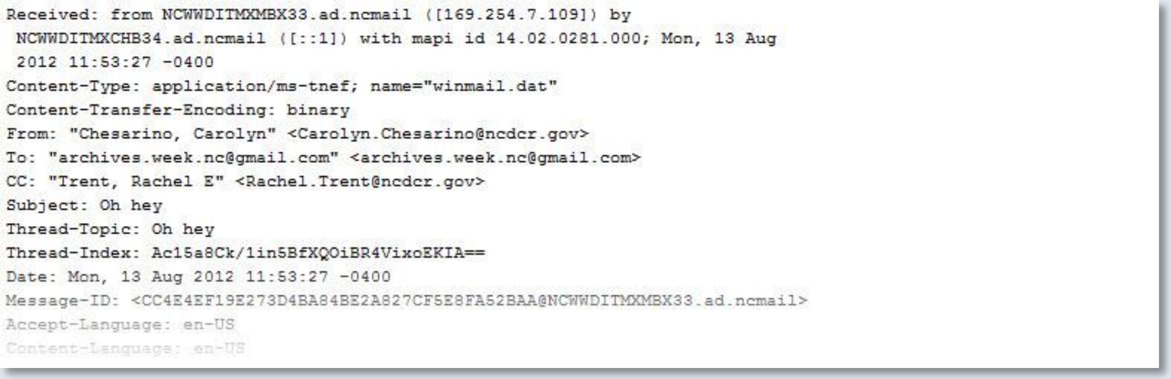
2. If the email message can only be opened in your email client, use a different format. Many email clients have developed their own proprietary format for email, and these should not be used when saving individual messages. Instead, the message should be saved in one of the general formats listed elsewhere in this document, such as plain text (see 2.2.1 Plain Text ) or PDF (see 2.1.1 PDF/A-1a). If the email is a plain text document, it should be able to be opened in a simple text editor, like Notepad or TextEdit. Plain text emails may carry the .txt extension, but they may also carry .eml or another extension. When selecting a file format in your email client, plain text formats may be identified under various titles, such as “EML,” “Email Message,” “Plain Text,” or “Show Original.”

⚠ **Note: If your file does not have the extension .txt, your computer may not know to open the file in a text editor. Test the file by manually opening it in Notepad, TextEdit, or another text editor on your computer.**

3. Email header information should be included in the file. The email “header” contains technical information that is very important in demonstrating the authenticity of an email during e-discovery or a public records request. Most important is the email address of the sender and recipient(s); many email file formats preserve the name of the sender and recipient(s), but not the email addresses. Be aware that email headers may be structured differently depending on the file format and the email client (see examples below). If you export messages from your email client as PDF, this metadata will probably not be included. You may need to export messages in a format that includes header metadata, and then convert that format to PDF.



```
MIME-Version: 1.0
Received: by 10.231.81.130 with HTTP; Mon, 13 Aug 2012 09:00:40 -0700 (PDT)
Date: Mon, 13 Aug 2012 12:00:40 -0400
Delivered-To: archives.week.nc@gmail.com
Message-ID: <CAAgFNi00hrU0DQfQPrZFk0C2T8q2b5JOPjmydvU7dG0SbFP8Tw@mail.gmail.com>
Subject: Important dates
From: Rachel Trent <archives.week.nc@gmail.com>
To: archives.week.nc@gmail.com
Content-Type: multipart/alternative; boundary=0003255756869f596404c727cbfb
```



```
Received: from NCWWDITMXMBX33.ad.ncmail ([169.254.7.109]) by
 NCWWDITMXCHB34.ad.ncmail ([::1]) with mapi id 14.02.0281.000; Mon, 13 Aug
 2012 11:53:27 -0400
Content-Type: application/ms-tnef; name="winmail.dat"
Content-Transfer-Encoding: binary
From: "Chesarino, Carolyn" <Carolyn.Chesarino@ncdcr.gov>
To: "archives.week.nc@gmail.com" <archives.week.nc@gmail.com>
CC: "Trent, Rachel E" <Rachel.Trent@ncdcr.gov>
Subject: Oh hey
Thread-Topic: Oh hey
Thread-Index: Ac15a8Ck/1in5BfXQOiBR4VixoEKIA==
Date: Mon, 13 Aug 2012 11:53:27 -0400
Message-ID: <CC4E4EF19E273D4BA84BE2A827CF5E8FA52BAA@NCWWDITMXMBX33.ad.ncmail>
Accept-Language: en-US
Content-Language: en-US
```

**Email headers appear differently depending on format and email client. Note that although the two partial emails header above are formatted differently, both contain important information like date, sender email address, recipient email address, and subject.**

## 2.12 Webpages

Websites are usually collections of numerous webpages that are intellectually related and meant to be explored as a whole. The North Carolina Department of Cultural Resources's site www.ncdcr.gov, for example, is a website that contains numerous webpages, including a, "About" page, a "News" web feed page, and a dynamic "Map" page. Each of these pages, in turn, is made up of multiple files of various file formats. A single webpage may include HTML, CSS, executable files, images, videos, audio, fonts, PDFs, and more. These complex digital entities, moreover, are often embedded in dense hyperlinked contexts, so that a single webpage removed from its context loses much of its meaning.

The goal of archiving a webpage is to collect all of the files, embedded content, and linked resources that originally made up the original webpage, and to be able to continue presenting the webpage as it originally appeared to visitors. Where possible, webpages should be collected in the context of the websites and linked webpages in which they were located at the time of capture. Several current web archiving services utilize the open-source Heritrix tool, which can perform large-scale archival web crawls and captures.[32] The Department of Cultural Resources performs large-scale captures of state government websites and social media content, utilizing the Web Archive (WARC) format and the Archive-It web archiving service. Currently, local government websites are not actively captured as a part of this program. Although parts of some local government sites are incidentally captured in the archive, local governments should not currently rely on this program to archive their websites.

### 2.12.1 Web Archive (.warc, .war)

Many web archiving services utilize the Web Archive, or WARC, format. This format specifies a standard for combining multiple digital resources into a single, aggregate file with descriptive information. The WARC format is an extension of the ARC File Format that has been used to store web crawls by the Internet Archive since 1996. Beginning in 2005, the Internet Archive developed the WARC format in consultation with the International Internet Preservation Consortium (IIPC) to extend and replace the ARC format. Published as ISO 28500:2009, WARC is an open, publicly documented standard.[33]

The WARC file format is designed to be used in the large-scale collection and bulk harvesting of web archives through tools built around the Heritrix open-source tool.[34] The State Archives and State Library of North Carolina currently crawl state government websites through

[32] For documentation of the Heritrix tool, see the Internet Archive's confluence page, https://webarchive.jira.com/wiki/display/Heritrix/Heritrix (accessed 5/29/2012).

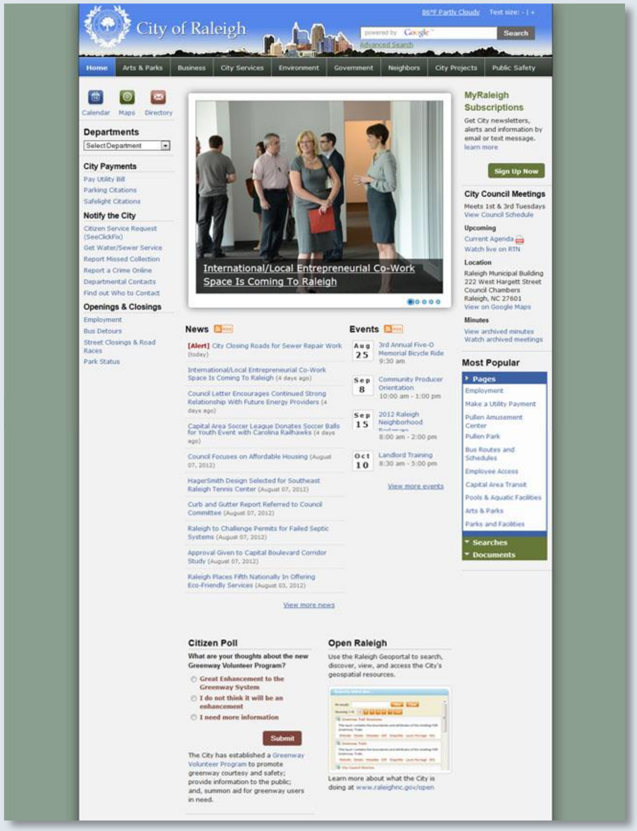
[33] ISO 28500:2009, Information and documentation -- WARC file format is available from ISO for purchase, http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44717 (accessed 5/15/2012). The draft standard that was the basis for approval, ISO/DIS 28500, is at http://bibnum.bnf.fr/WARC/warc_ISO_DIS_28500.pdf (accessed 5/15/2012).

[34] See documentation of the Heritrix tool at the Internet Archive: https://webarchive.jira.com/wiki/display/Heritrix/Heritrix.

Archive-It. The archived websites can be accessed online at www.webarchives.ncdcr.gov. To submit a new website to the web archive, please contact webarchives@ncdcr.gov.

### 2.12.2 PDF/A-1a (.pdf) (ISO 19005-1 compliant PDF/A)

See also 2.1.1 PDF/A-1a (.pdf) (ISO 19005-1 compliant PDF/A). It is recommended that local governments use vendor services that archive websites in WARC or other formats that allow websites to be rendered as they were originally created, with active hyperlinks and other interactive material. However, where this is not possible, local governments may consider capturing websites in the PDF format. A special type of PDF format, PDF/A-1a, is specially designed to preserve files long-term, and the State Archives recommends that PDF/A-1a be used wherever possible. More information on this format can be found in section 2.1.1 PDF/A-1a (.pdf) (ISO 19005-1 compliant PDF/A).



**Departments**

**City Payments**
Pay Utility Bill
Parking Citations
Safelight Citations

**Notify the City**
Citizen Service Request (SeeClickFix)
Get Water/Sewer Service
Report Missed Collection
Report a Crime Online
Departmental Contacts
Find out Who to Contact

**Openings & Closings**
Employment
Bus Detours
Street Closings & Road Races
Park Status

International/Local Entrepreneurial Co-Work Space Is Coming To Raleigh

**News**

[Alert] City Closing Roads for Sewer Repair Work (today)
International/Local Entrepreneurial Co-Work Space Is Coming To Raleigh (4 days ago)
Council Letter Encourages Continued Strong Relationship With Future Energy Providers (4 days ago)
Capital Area Soccer League Donates Soccer Balls for Youth Event with Carolina Railhawks (4 days ago)
Council Focuses on Affordable Housing (August 07, 2012)
HagerSmith Design Selected for Southeast Raleigh Tennis Center (August 07, 2012)
Curb and Gutter Report Referred to Council Committee (August 07, 2012)
Raleigh to Challenge Permits for Failed Septic Systems (August 07, 2012)
Approval Given to Capital Boulevard Corridor Study (August 07, 2012)
Raleigh Places Fifth Nationally In Offering Eco-Friendly Services (August 03, 2012)
View more news

**Events**

Aug 25 3rd Annual Five-O Memorial Bicycle Ride 9:30 am
Sep 8 Community Producer Orientation 10:00 am - 1:00 pm
Sep 15 2012 Raleigh Neighborhood Exchange 8:00 am - 2:00 pm
Oct 10 Landlord Training 8:30 am - 5:00 pm
View more events

**MyRaleigh Subscriptions**
Get City newsletters, alerts and information by email or text message.
learn more

**City Council Meetings**
Meets 1st & 3rd Tuesdays
View Council Schedule

**Upcoming**
Current Agenda
Watch live on RTN

**Location**
Raleigh Municipal Building
222 West Hargett Street
Council Chambers
Raleigh, NC 27601
View on Google Maps

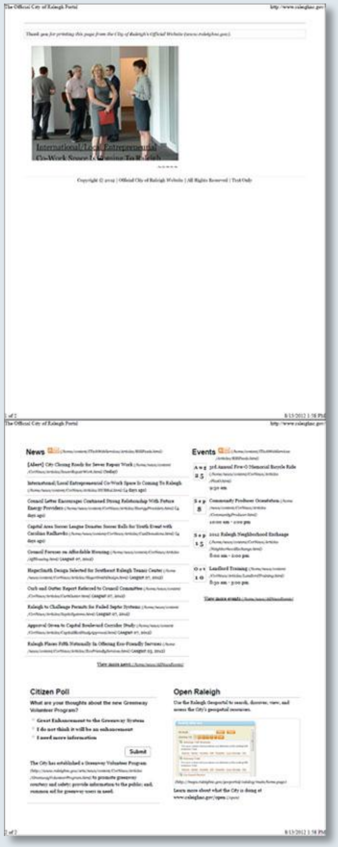
**Minutes**
View archived minutes
Watch archived meetings

**Most Popular**
Pages
Employment
Make a Utility Payment
Pullen Amusement Center
Pullen Park
Bus Routes and Schedules
Employee Access
Capital Area Transit
Pools & Aquatic Facilities
Arts & Parks
Parks and Facilities
Searches
Documents

**Citizen Poll**
What are your thoughts about the new Greenway Volunteer Program?
○ Great Enhancement to the Greenway System
○ I do not think it will be an enhancement
○ I need more information

The City has established a Greenway Volunteer Program to promote greenway courtesy and safety; provide information to the public; and, summon aid for greenway users in need.

**Open Raleigh**
Use the Raleigh Geoportal to search, discover, view, and access the City's geospatial resources.
Learn more about what the City is doing at www.raleighnc.gov/open


The Official City of Raleigh Portal http://www.raleighnc.gov/

*Thank you for printing this page from the City of Raleigh's Official Website (www.raleighnc.gov).*

International/Local Entrepreneurial Co-Work Space Is Coming To Raleigh

Copyright © 2012 | Official City of Raleigh Website | All Rights Reserved | Text Only

1 of 2 8/13/2012 1:58 PM

The Official City of Raleigh Portal http://www.raleighnc.gov/

**News** [illegible]

[Alert] City Closing Roads for Sewer Repair Work [illegible] (today)
International/Local Entrepreneurial Co-Work Space Is Coming To Raleigh [illegible] (4 days ago)
Council Letter Encourages Continued Strong Relationship With Future Energy Providers [illegible] (4 days ago)
Capital Area Soccer League Donates Soccer Balls for Youth Event with Carolina Railhawks [illegible] (4 days ago)
Council Focuses on Affordable Housing [illegible] (August 07, 2012)
HagerSmith Design Selected for Southeast Raleigh Tennis Center [illegible] (August 07, 2012)
Curb and Gutter Report Referred to Council Committee [illegible] (August 07, 2012)
Raleigh to Challenge Permits for Failed Septic Systems [illegible] (August 07, 2012)
Approval Given to Capital Boulevard Corridor Study [illegible] (August 07, 2012)
Raleigh Places Fifth Nationally In Offering Eco-Friendly Services [illegible] (August 03, 2012)
View more news [illegible]

**Events** [illegible]

Aug 25 3rd Annual Five-O Memorial Bicycle Ride [illegible] 9:30 am
Sep 8 Community Producer Orientation [illegible] 10:00 am - 1:00 pm
Sep 15 2012 Raleigh Neighborhood Exchange [illegible] 8:00 am - 2:00 pm
Oct 10 Landlord Training [illegible] 8:30 am - 5:00 pm
View more events [illegible]

**Citizen Poll**
What are your thoughts about the new Greenway Volunteer Program?
○ Great Enhancement to the Greenway System
○ I do not think it will be an enhancement
○ I need more information
Submit
The City has established a Greenway Volunteer Program [illegible] to promote greenway courtesy and safety; provide information to the public; and, summon aid for greenway users in need.

**Open Raleigh**
Use the Raleigh Geoportal to search, discover, view, and access the City's geospatial resources.
[illegible]
Learn more about what the City is doing at www.raleighnc.gov/open [illegible]

2 of 2 8/13/2012 1:58 PM

⚠ PDF is designed to replicate the "look" of documents—not the interactivity of websites. Many web browsers, such as Firefox®, Internet Explorer®, and Google Chrome™, have the option to "save as" or print to PDF. Depending on how your website is built, this option may drastically change the appearance of the website. Interactive elements, including hyperlinks, videos, sound, forms, and scripts will lose their functionality. Note the example page above. On the left is the original website captured with a screenshot. On the right is the same website printed to PDF using Firefox®.

## 2.13 Geospatial Vector Datasets

**Note: Extensive documentation covering proper preparation of geospatial data and current North Carolina geospatial preservation standards and workflows can be found at** www.geomapp.net**. Especially relevant is the** Geospatial Data File Formats Reference Guide **(July 2011).**

Currently, the State Archives of North Carolina collects statewide geospatial data from NC OneMap, a clearinghouse for North Carolina geospatial resources. The archival entity collected by the State Archives is the shapefile. The shapefile format, formally known as the ESRI Shapefile Format, is an open specification designed by Environmental Systems Research Institute, Inc (Esri®) for the transfer of data between Esri® and non-Esri® products. The format is defined in *ESRI Shapefile Technical Description: An ESRI White Paper—July 1998.*[35]

Shapefiles store nontopological geometry and attribute information for the spatial features of a data set. The geometry for a feature is stored as a shape comprising a set of vector coordinates. A single shapefile is, in fact, a collection of several distinct files. At a minimum, the shapefile consists of a main file (.shp) an index file (.shx), and a database or "dBASE" file (.dbf). The main file contains a record of each point, line, and area in the shapefile, with each record being described by a list of its vertices. The index file lists the location of each record in the main file, and the database (or "dBASE") file contains the attributes of each record. Shapefiles typically have several additional, optional component files.[36] Shapefiles collected by the State Archives of North Carolina contain the following seven component files:

- **.shp** — Main file: direct access, variable-record-length file in which each record describes a shape with a list of its vertices
- **.shx** — Index file: list of records containing the offset of the corresponding main file record from the beginning of the main file
- **.dbf** — dBASE file: table containing feature attributes with one record per feature and a one-to-one relationship between geometry and attributes based on record number
- **.prj** — Projection Definition file: coordinate system information

[35] "ESRI Shapefile Technical Description: An ESRI White Paper—July 1998," Environmental Systems Research Institute, Inc. http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf (accessed 5/15/2012).

[36] "ESRI Shapefile Technical Description: An ESRI White Paper—July 1998," Environmental Systems Research Institute, Inc. http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf (accessed 5/15/2012); Library of Congress, "ESRI Shapefile," *Sustainability of Digital Formats: Planning for Library of Congress Collections*, http://digitalpreservation.gov/formats/fdd/fdd000280.shtml (accessed 5/16/2012).

- **.sbn** — Part 1 of spatial index for read-write instances of the Shapefile format: if present, essential for correct processing
- **.sbx** — Part 2 of spatial index for read-write instances of the Shapefile format: if present, essential for correct processing
- **.shp.xml** — Geospatial metadata file: metadata in XML format following either the Federal Geographic Data Committee's (FGDC) Content Standard for Geospatial Metadata (CSDGM) FGDC-STD-001-1998 or ISO 19115:2003, with the optional addition of the Esri® Metadata Profile[37]

[37] Federal Geographic Data Committee's (FGDC), "Content Standard for Digital Geospatial Metadata" http://www.fgdc.gov/metadata/csdgm/ (accessed 5/16/2012); ISO 19115:2003 "Geographic information – Metadata" http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020 (accessed 5/16/2012); FGDC, "Draft North American Profile of ISO19115:2003 - Geographic information – Metadata" http://www.fgdc.gov/standards/projects/incits-l1-standards-projects/NAP-Metadata/napMetadataProfileV11_7-26-07.pdf/view (accessed 5/16/2012); "Esri Profile of the *Content Standard for Digital Geospatial Metadata*," 2003 http://www.esri.com/metadata/esriprof80.html (accessed 5/16/2012).